Unterrichtswissenschaft (2023) 51:455–478
https://doi.org/10.1007/s42010-023-00177-y

ALLGEMEINER TEIL

Validierung eines Tests zur Erfassung des mathematikdidaktischen Wissens von Lehramtsstudierenden der Primarstufe

Lars Jenßen · Malte Lehmann · Christin Laschke · Bettina Roesken-Winter · Katja Eilerts

Eingegangen: 8. November 2022 / Überarbeitet: 21. Juni 2023 / Angenommen: 30. Juni 2023 / Online publiziert: 18. Juli 2023

Zusammenfassung Mathematikdidaktisches Wissen stellt eine zentrale Komponente der professionellen Kompetenz von (angehenden) Primarstufenlehrkräften im Bereich Mathematik dar, die mit situationsspezifischen Fertigkeiten zusammenhängt und, vermittelt über das professionelle Handeln, Effekte auf die mathematischen Leistungen von Schüler:innen haben kann. Um mathematikdidaktisches Wissen im Zusammenhang mit anderen Komponenten professioneller Kompetenz (z. B. Emotionen, Überzeugungen, Handlungsplanung, Instruktionsqualität) untersuchen zu können, bedarf es eines ökonomischen und frei verfügbaren Tests, welcher gängige Gütekriterien hinreichend erfüllt. Der vorliegende Beitrag stellt die Validierung eines solchen Tests vor. Entsprechend der Teststandards werden Grundnahmen in den Evidenzkategorien Inhalt, Struktur und Beziehungen zu anderen Variablen un-

✉ Lars Jenßen
Institut für Erziehungswissenschaften, Erziehungswissenschaftliche Methodenlehre, Humboldt-Universität zu Berlin, Berlin, Deutschland
E-Mail: lars.jenssen@hu-berlin.de

Malte Lehmann · Bettina Roesken-Winter · Katja Eilerts
Institut für Erziehungswissenschaften, Mathematik und ihre Didaktik der Primarstufe, Humboldt-Universität zu Berlin, Berlin, Deutschland

Malte Lehmann
E-Mail: malte.lehmann@hu-berlin.de

Bettina Roesken-Winter
E-Mail: bettina.roesken-winter@hu-berlin.de

Katja Eilerts
E-Mail: katja.eilerts@hu-berlin.de

Christin Laschke
Fachbezogener Erkenntnistransfer, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel, Deutschland
E-Mail: laschke@leibniz-ipn.de

tersucht, um zu prüfen, ob die Testscores Schlussfolgerungen zum Konstrukt mathematikdidaktisches Wissen von Lehramtsstudierenden der Primarstufe zulassen. Die Ergebnisse liefern Validitätsargumente, die auf eine hinreichend hohe Reliabilität des Tests hinweisen und für theoriekonforme Schlussfolgerungen basierend auf den Testwerten sprechen. Der Beitrag schließt ab mit einer integrativen Betrachtung der Validierungsevidenzen, die für den Test bisher vorliegen.

Schlüsselwörter Primarstufe · Lehramtsstudierende · Mathematikdidaktisches Wissen · Testentwicklung · Validierung

Validating an assessment of pre-service primary teachers' mathematics pedagogical content knowledge

Abstract Mathematics pedagogical content knowledge is a central component of the professional competence of pre-service primary school teachers in the field of mathematics, which is related to situation-specific skills and can have effects on students' performance in mathematics. In order to be able to examine mathematic pedagogical content in relation to other components of professional competence (e.g., emotions, beliefs, action planning, instructional quality), an economical and freely available test that sufficiently fulfills common quality criteria is needed. This paper presents the validation of such a test. According to the test standards and in order to proof whether test scores allow conclusions on pre-service primary school teachers' mathematics pedagogical content knowledge, key assumptions according to content, structure and relationships to other variables were examined. The results provide validity arguments, indicating a sufficiently high reliability of the test and support theory-conform conclusions based on the test scores. The paper concludes with an integrative review of the validation evidence available for the test to date.

Keywords Primary school · Pre-service teachers · Mathematics pedagogical content knowledge · Test development · Validation

1 Einleitung

Die Untersuchung der professionellen Kompetenz von Lehrkräften als auch von Lehramtsstudierenden war in den letzten 20 Jahren geprägt durch umfangreiche empirische Forschungsprojekte und stellt nach wie vor einen wesentlichen Schwerpunkt in der Schul- und Unterrichtsforschung dar. Um dieses Forschungsfeld weiter voranzutreiben, werden valide und reliable Testverfahren zur Erfassung von Facetten professioneller Kompetenz benötigt, insbesondere für die Messung des mathematikdidaktischen Wissens mittels eines ökonomisch einsetzbaren Instruments.

In neueren Konzeptionen wird das professionelle Wissen von Lehrkräften als Bestandteil der professionellen Kompetenz angesehen, welches neben affektiv-motivationalen Dispositionen vermittelt über situationsspezifische Fertigkeiten das professionelle Handeln (*Performanz*) von Lehrkräften bedingt (Blömeke et al. 2015a). Fachdidaktisches Wissen stellt neben fachwissenschaftlichem und allgemein-päd-

agogischem Wissen eine wesentliche Facette professionellen Wissens von Lehrkräften dar (Baumert und Kunter 2006; Shulman 1986). Das fachdidaktische Wissen kann je nach Domäne (z. B. Mathematik) weiter ausdifferenziert werden (Depaepe et al. 2013; Neumann et al. 2019). Im Bereich Mathematik wird das fachdidaktische Wissen als mathematikdidaktisches Wissen (in der *scientific community* mit MPCK für *mathematics pedagogical content knowledge* abgekürzt) bezeichnet (Ball et al. 2008). Dieses umfasst u. a. das Lehren von Mathematik, das als gemeinsame Basis aller Inhaltsbereiche verstanden wird und im Gegensatz zum fachlichen Wissen, das mathematisches Wissen in Breite und Tiefe umfasst (MCK als Abkürzung für *mathematics content knowledge*: Ball et al. 2008) und eher interaktionsbezogene Aspekte beinhaltet (Carpenter et al. 1989).

MPCK wird in größeren Studien üblicherweise in Form von Tests erfasst und in Studien mit kleinerer Stichprobe teilweise eher über qualitative Forschungsansätze erhoben (Depaepe et al. 2013). Tests zur Erfassung von MPCK bestehen beispielsweise für Sekundarstufenlehrkräfte (z. B. COACTIV: Krauss et al. 2008a) oder angehende frühpädagogische Fachkräfte (z. B. KomMa-MPCK-Test: Blömeke et al. 2015b). Ein etablierter Test zur Erfassung von MPCK von Primarstufenlehrkräften wurde unter Berücksichtigung nationaler Besonderheiten der U.S.A entwickelt (Hill et al. 2008) und hat somit eine stark limitierte Aussagekraft für Lehramtsstudierende für die Primarstufe in Deutschland, deren Lerngelegenheiten sich hinsichtlich Umfang und Inhalt von denen in den U.S.A unterscheiden (König et al. 2010). Der Ein weiterer Test zur Erfassung von MPCK speziell für Lehramtsstudierende für die Primarstufe wurde im Rahmen der Large-Scale-Studie TEDS-M (*Teacher Education and Development Study: Learning to Teach Mathematics)* entwickelt und umfasst 32 Items in unterschiedlichen Formaten (Multiple-Choice, Complex-Multiple-Choice, offenes Antwortformat). Die Skalierung des Tests erfolgte vor dem Hintergrund der Item-Response-Theorie (Blömeke et al. 2010a), die immer eine größere Stichprobe für die komplexen Modellierungen erfordern (Rost 2004). Die Besonderheit des Tests besteht vor allem darin, dass Gemeinsamkeiten der Curricula aus insgesamt 15 teilnehmenden Ländern als inhaltliche Grundlage für die Itementwicklung genutzt wurden und daher nationale Besonderheiten nicht abgebildet sind (Blömeke et al. 2010b). Zudem ist der Test nicht frei verfügbar.

Insgesamt lässt sich feststellen, dass ein Mangel an frei verfügbaren Tests zur Erfassung von MPCK von Lehramtsstudierenden für die Primarstufe besteht, welche die Inhalte des Lehramtsstudiums in Deutschland abbilden. Vor allem besteht ein Mangel an solchen Tests, die ökonomisch einsetzbar sowie reliabel und valide sind. Diese werden jedoch benötigt, um Zusammenhänge mit anderen Bestandteilen professioneller Kompetenz wie Beliefs oder situationsspezifischen Fertigkeiten abzubilden. Ebenso fehlen ökonomisch einsetzbare Tests für Forschung im Kontext von Professionalisierungsprozessen in der Lehrkräftebildung, so dass zusätzlich zu gegenstandsspezifischen Aspekten das fachdidaktische Wissen mit kurzer Testbearbeitungszeit erfasst wird und Akzeptanzproblemen, bedingt durch knappe Zeitressourcen, vorgebeugt werden kann.

Der vorliegende Beitrag stellt einen Test zur Erfassung des mathematikdidaktischen Wissens von Lehramtsstudierenden der Primarstufe vor und beschreibt das Vorgehen bei der Validierung von Testwertinterpretationen, die der Test ermöglichen

soll. Zunächst wird das Konstrukt MPCK hinsichtlich des Inhalts, der Struktur, der Erlernbarkeit und Zusammenhängen zu anderen kognitiven Konstrukten theoretisch aufgearbeitet. Im Anschluss werden Validierungsaspekte in Bezug auf Tests zur Erfassung des professionellen Wissens von angehenden pädagogischen Fachkräften zusammengestellt. Danach wird die Entwicklung des Tests zur Erfassung des mathematikdidaktischen Wissens von Lehramtsstudierenden der Primarstufe vorgestellt. Ausgehend davon werden die Ziele und die zu prüfenden Grundannahmen im Rahmen des Validierungsprozesses vorgestellt und die dazu gewonnenen Evidenzen abschließend integrierend betrachtet und diskutiert.

2 Mathematikdidaktisches Wissen von Lehramtsstudierenden der Primarstufe

Mit den grundlegenden Arbeiten von Shulman (1986, 1987) wurde das fachdidaktische Wissen als eine zentrale Komponente der kognitiven Dispositionen des professionellen Wissens von Lehrkräften eingeführt. Darauf aufbauend existiert eine Vielzahl unterschiedlicher Definitionen fachdidaktischen Wissens, die jeweils andere Schwerpunkte setzen. So konnten Depaepe et al. (2013) in ihrem Systematic Review mehrere Komponenten von MPCK identifizieren, die Wissensbereiche zu (Fehl-)Vorstellungen von Schüler:innen, zu didaktischen Methoden, mathematischen Aufgaben und kognitiven Anforderungen, curriculares Wissen, Kontextwissen sowie inhaltliches und pädagogisches Wissen umfassen. Dabei konnten Depaepe et al. (2013) aufzeigen, dass es im Wesentlichen zwei Perspektiven auf MPCK gibt: eine kognitive, welche eher ein Wissen über das Unterrichten beschreibt, das unabhängig vom Unterrichtskontext erlernt und angewendet werden kann und eine situative Perspektive, die Handlungswissen fokussiert, das im Kontext des Mathematikunterrichts situiert ist.

Eine im deutschsprachigen Raum etablierte Konzeptualisierung des fachdidaktischen Wissens wurde im *COACTIV* Projekt (Baumert et al. 2010) vorgenommen. Hier wird dieses als „ein besonders unterrichts- und schülerbezogenes mathematisches Wissen und Können“ beschrieben (Baumert und Kunter 2006, S. 495), das sich in Subdimensionen unterteilt. Diese Subdimensionen umfassen im Wesentlichen Wissen über das didaktische und diagnostische Potenzial von Aufgaben, Wissen über die kognitiven Anforderungen und impliziten Wissensvoraussetzungen von Aufgaben, ihre didaktische Sequenzierung und die langfristige curriculare Anordnung von Wissensinhalten, aber auch Wissen über Schüler:innenvorstellungen (Fehlkonzeptionen, typische Fehler, Strategien) und Diagnostik von Schüler:innenwissen und Verständnisprozessen sowie methodisches Wissen. Vergleichbar strukturieren Ball et al. (2008) das Konstrukt in ihrem *MKT-Framework* in die Subdimensionen *Knowledge of Content and Students, Knowledge of Content and Teaching* und *Knowledge of Content and Curriculum*. Buchholtz et al. (2014) weisen darauf hin, dass die Konzeptualisierungen häufig nah an fachmathematischen Aspekten vorgenommen werden und andere Aspekte wie fachdidaktisches Wissen zu Lehr-Lern-Prozessen weniger Berücksichtigung erfahren. Deswegen wurden im Projekt TEDS-LT auch

Fragen zu mathematischen Curricula und Bildungsstandards, aber auch zur psychologischen Beschreibung von mathematischen Denkhandlungen integriert.

Eine Zusammenschau dieser etablierten inhaltlichen Ausdifferenzierungen des mathematikdidaktischen Wissens von Lehramtsstudierenden für die Primarstufe, welche nationale Spezifika berücksichtigt und einzelne Bereiche entsprechend der Lehramtsausbildung akzentuiert, ergibt eine Konzeptualisierung von vier Inhaltsbereichen von MPCK:

1. Lernwege und Verständnis von Schüler:innen: Dieser Inhaltsbereich umfasst Wissen zu (Fehl-)Vorstellungen von Schüler:innen zu mathematischen Inhalten und Prozessen der Primarstufe – einschließlich der Klassenstufen 5 und 6[1] (z. B. Zahlaspekte natürlicher Zahlen und Grundvorstellungen zur Subtraktion)
2. Curricula und Lernkontexte: Dieser Inhaltsbereich umfasst Wissen zu Leitideen und Kompetenzen, wie sie in den Bildungsstandards für die Primarstufe (Kultusministerkonferenz 2022) formuliert sind (z. B. Wissen zu primarstufenspezifischen Inhalten der Leitidee *Größen und Messen*)
3. Aufgaben und Material: Dieser Inhaltsbereich umfasst vor allem Wissen zu spezifischen Aufgabentypen und Materialien, die als relevant für die Primarstufe erachtet werden (z. B. Kennzeichen von Aufgabenformaten, die produktives Üben ermöglichen)
4. Lehren und Unterrichten: Dieser Inhaltsbereich umfasst Wissen zu Wegen der Vermittlung mathematischer Inhalte für die Primarstufe (z. B. Repräsentationen zum Bündeln und Entbündeln im Dezimalsystem oder die Anwendung des EIS-Prinzips nach Bruner).

Diese hier vorgestellte Konzeptualisierung soll dabei weniger als ein theoretisches Modell verstanden werden, sondern im Sinne der Anschlussfähigkeit als eine Heuristik zur Deskription des mathematikdidaktischen Wissens von Lehramtsstudierenden für die Primarstufe. Die Aufgabenbeispiele in den Abb. 1, 2 und 3 sollen die Umsetzung der Konzeptualisierung für die Testkonstruktion verdeutlichen.

Auf struktureller Ebene lässt sich neben der zuvor dargestellten inhaltlichen Ausdifferenzierung eine hohe positive Assoziation mit dem MCK feststellen. Graeber und Tirosh (2008) weisen darauf hin, dass früher kein differenziertes Verständnis von MCK und MPCK existierte, was eine entsprechende Erfassung erschwerte. Ausgehend von Shulman (1986) lassen sich jedoch beide Wissensfacetten sowohl theoretisch als auch empirisch trennen. Für angehende Primarstufenlehrkräfte in Deutschland zeigten Analysen im Rahmen des Projekts TEDS-M eine manifeste Korrelation von $r = 0{,}62$ zwischen MPCK und MCK (Blömeke et al. 2010a). Auf der Konstruktebene wird MCK meist als Voraussetzung für die Entwicklung von MPCK konzeptualisiert (Baumert et al. 2010), was auch von neueren Studien empirisch gestützt wird (Agathangelou und Charalambous 2020). In querschnittlichen Studien wird dennoch üblicherweise eine Korrelation angenommen (Blömeke et al. 2010a), da davon ausgegangen werden kann, dass sich MPCK und MCK glei-

[1] In Berlin und Brandenburg umfasst die Primarstufe auch die Klassenstufen 5 und 6. In allen anderen Bundesländern nicht.

chermaßen im Verlauf ihres Erwerbs im Lehramtsstudium wechselseitig bedingen bzw. miteinander zusammenhängen, ohne, dass eine Kausalannahme getroffen wird (Krauss et al. 2008b). Es kann davon ausgegangen werden, dass MCK und MPCK mit voranschreitender Professionalisierung zunehmend miteinander vernetzt werden und entsprechend repräsentiert sind (Liljedahl et al. 2009).

Neben den Inhalten und deren Strukturierung können zur Beschreibung des professionellen Wissens weitere Merkmale wie die Erlernbarkeit und die Abgrenzung von allgemein-kognitiven Fähigkeiten herangezogen werden (Blömeke et al. 2015a). Diese werden auch bei der Konzeptualisierung von Koeppen et al. (2008) für Kompetenzen von Schüler:innen benannt und sind auch für Lehramtsstudierende anzunehmen. Die Erlernbarkeit betonte bereits Weinert (2001) als zentrales Merkmal von Kompetenz und damit auch als wesentlichen Unterschied zu allgemein-kognitiven Fähigkeiten. Blömeke et al. (2015a) zufolge wird professionelles Wissen im Rahmen von Aus- und Weiterbildung erworben. Dementsprechend nimmt mit voranschreitender Professionalisierung im Rahmen der Lehramtsausbildung auch das professionelle Wissen der Lehramtsstudierenden zu (Berliner 2001; Krauss et al. 2008a). Auch für MPCK kann daher angenommen werden, dass dieses z. B. im Rahmen von Vorlesungen und Seminaren im Studium des Primarstufenlehramts erworben wird.

Mit Blick auf die Abgrenzbarkeit zu allgemein-kognitiven Fähigkeiten hat sich gezeigt, dass domänenspezifisches Wissen und damit auch die erbrachte Leistung in dieser Domäne von Intelligenz empirisch trennbar ist, auch wenn diese den Erwerb des Wissens und die Leistungserbringung positiv bedingt (Thorsen et al. 2014). Rindermann (2007) zeigte in seinen Analysen zu Schulleistungsstudien, dass Leistungen über die Fächer hinweg stark miteinander zusammenhängen, was er als Argument dafür sah, dass diese Studien lediglich allgemein-kognitive Fähigkeiten erfassen würden. Tatsächlich ist aber davon auszugehen, dass Leistungen in verschiedenen Bereichen deswegen miteinander zusammenhängen, weil sie alle gleichermaßen kognitive Merkmale darstellen (Prenzel et al. 2007). Studien mit angehenden frühpädagogischen Fachkräften zeigten ebenso, dass Wissensbereiche miteinander korrelieren (Blömeke, Jenßen et al. 2015) und dass diese Zusammenhänge über den Einfluss allgemein-kognitiver Fähigkeiten hinaus gehen (Blömeke et al. 2022; Jenßen et al. 2019). Studien mit Referendar:innen stützen diese Befunde auch für angehende Lehrkräfte (Voss et al. 2011).

Werden Messverfahren zur Erfassung von MPCK entwickelt, müssen also Aspekte wie die Inhalte und die Strukturierung derer sowie die Beziehung zu anderen Variablen geprüft werden, um Hinweise auf die Validität zu gewinnen. Im Folgenden werden mögliche Validierungsschritte und deren Umsetzung in bisherigen Studien aufgezeigt.

3 Validierung von Testverfahren zur Erfassung des professionellen Wissens von angehenden Lehrkräften

Das Verständnis von Validierung hat sich in den letzten Jahren gewandelt, was sich vor allem in der Kompetenzforschung zeigt (Jenßen et al. 2015). Demnach ist Va-

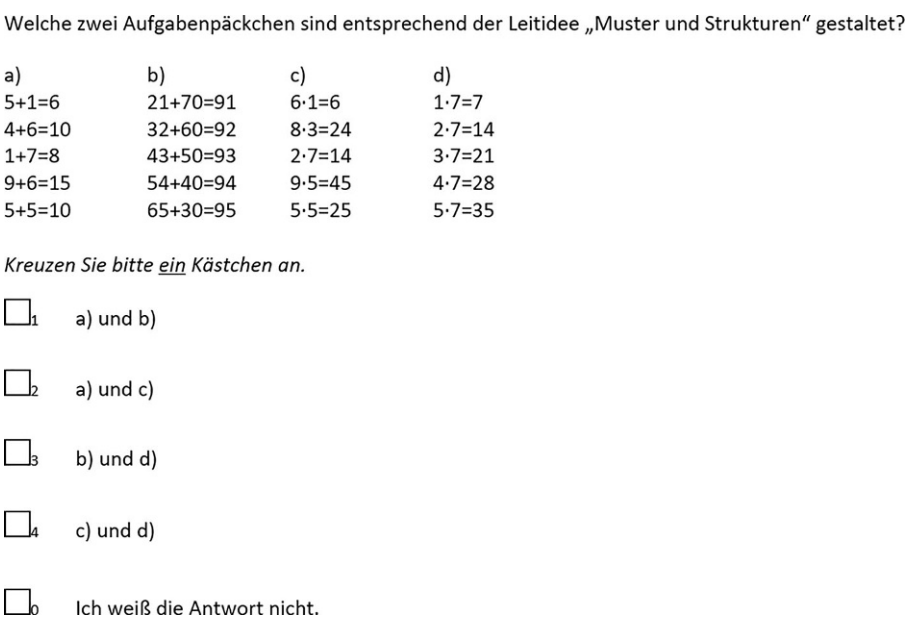
lidität nicht mehr als eine Eigenschaft eines Messverfahrens zu verstehen, vielmehr steht die Überprüfung von Schlussfolgerungen (Testwertinterpretationen), die aus den Testscores gezogen werden sollen, im Vordergrund (Kane 2013). Nach Kane (2001) kann zwischen mehreren Arten von Testwertinterpretationen wie der *Bewertung, Verallgemeinerung, Extrapolation, Erklärung* und der *Entscheidungsfindung* unterschieden werden. Die angestrebte *erklärende Inferenz* liegt beispielsweise dann vor, wenn Testscores als Indikatoren für ein theoretisches Konstrukt interpretiert werden können. Aus den Testwertinterpretationen lassen sich Grundannahmen ableiten, die empirisch zu prüfen sind (Hartig et al. 2020). Dafür wurden in den von AERA, APA und NCME (2014) gemeinsam entwickelten *Standards for Educational and Psychological Testing* Evidenzkategorien beschrieben, anhand derer empirische Argumente zur Stützung von validen Schlussfolgerungen gesammelt werden können. Diese betreffen den *Inhalt* des Messverfahrens, die *Struktur* des Messverfahrens (v. a. Dimensionalität bzw. Faktorstruktur), *Antwortprozesse*, die der Bearbeitung von Testaufgaben zugrunde liegen, und *Beziehungen zu anderen Variablen*. Hartig et al. (2020) weisen darauf hin, dass bei Grundannahmen, die durch Evidenzen bezüglich der Beziehungen zu anderen Variablen geprüft werden, Aussagen zum erwarteten Bereich von Zusammenhängen (z. B. bei korrelativen Beziehungen) getroffen werden sollten. Dies ermöglicht, dass das Falsifikationsprinzip der empirischen Forschung auch bei der Validierung berücksichtigt wird. Als Fazit einer Validierung lässt sich „an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores“ (Messick 1992, p. 1487) ziehen.

Entsprechend des modernen Verständnisses von Validität konzentrieren sich Validierungsbemühungen zu Tests zur Erfassung des professionellen Wissens von Lehrkräften auf die zentralen Testwertinterpretationen, die mit dem Test ermöglicht werden sollen. Validierungen sowohl im COACTIV Projekt als auch bei TEDS-M fokussierten zunächst die *erklärende Inferenz* als angestrebte Testwertinterpretation. Im Vordergrund standen daher Validierungsbemühungen, die zeigten, dass die Testwerte als Indikatoren des theoretischen Konstrukts des professionellen Wissens angenommen werden können. Dementsprechend wurden Evidenzen zur Stützung dieser Testwertinterpretation gesammelt. Beispielhaft seien hier dimensionale Analysen zwischen MCK und MPCK von Lehramtsstudierenden für die Primarstufe bei TEDS-M (Blömeke et al. 2010a) oder die Untersuchung der Erlernbarkeit von MCK und MPCK von Lehramtsstudierenden für die Sekundarstufe der im Rahmen des Projekts COACTIV formulierten *Growing Knowledge Hypothesis* (Krauss et al. 2008a) genannt. Da, wie bereits weiter oben ausgeführt, professionelles Wissen als ein kognitives Konstrukt mit anderen allgemein-kognitiven und domänenspezifisch-kognitiven Konstrukten zusammenhängt, werden zur Validierung entsprechender Testverfahren häufig Schulnoten eingesetzt (z. B. bei angehenden frühpädagogischen Fachkräften: Blömeke et al. 2015b). Validierte Testverfahren zur Erfassung von MPCK von Lehramtsstudierenden für die Primarstufe sind bislang nicht verfügbar.

4 Entwicklung des Tests zum mathematikdidaktischen Wissen

Zur Schließung des Desiderats ist die Konstruktion eines Tests notwendig, welcher das MPCK von Lehramtsstudierenden für die Primarstufe ökonomisch, reliabel und valide erfasst und frei verfügbar ist. Um diesen Ansprüchen an das Testverfahren zu genügen, wurde entschieden, sich an der kognitiven und eher statischen Perspektive auf MPCK zu orientieren (Depaepe et al. 2013), vergleichbar zum COACTIV Projekt (Baumert et al. 2010) oder Ball et al. (2008). Dabei berücksichtigt der Fokus auf die kognitive Perspektive auf MPCK auch, dass der Test für Studierende in den ersten Studienjahren konzipiert wird. In diesem Zeitraum steht eher theoretisches Wissen über unterrichtsrelevante Aspekte im Vordergrund, das weniger situiert ist. Entsprechend sehen sich Studierende im Bachelor auch zunächst als Lernende und erst später im Masterstudium zunehmend als prospektiv Lehrende (Chen 2017). Die Autor:innengruppe entwickelte daher vor dem Hintergrund des interdisziplinären Projektteams entlang der oben vorgestellten Heuristik der vier Inhaltsbereiche (Lernwege und Verständnis von Schüler:innen, Curricula und Lernkontexte, Aufgaben und Material, Lehren und Unterrichten) Testaufgaben, die sich zur Erfassung des MPCK von Lehramtsstudierenden für die Primarstufe eignen und gemeinsam einen Leistungstest bilden sollen.

Abb. 1 zeigt ein Item aus dem Inhaltsbereich *Aufgaben und Material*. In der Leitidee *Muster und Strukturen* wird als Lerninhalt für die Schüler:innen der Primarstufe formuliert, dass sie Strukturen in arithmetischen Mustern erkennen und beschreiben können (Kultusministerkonferenz 2022). In der dritten Antwortmöglichkeit (b und d) werden Päckchen angesprochen, in denen Muster gefunden wer-


Welche zwei Aufgabenpäckchen sind entsprechend der Leitidee „Muster und Strukturen“ gestaltet?

a)	b)	c)	d)
5+1=6	21+70=91	6·1=6	1·7=7
4+6=10	32+60=92	8·3=24	2·7=14
1+7=8	43+50=93	2·7=14	3·7=21
9+6=15	54+40=94	9·5=45	4·7=28
5+5=10	65+30=95	5·5=25	5·7=35

Kreuzen Sie bitte ein Kästchen an.

☐1 a) und b)

☐2 a) und c)

☐3 b) und d)

☐4 c) und d)

☐0 Ich weiß die Antwort nicht.

Abb. 1 Beispielitem aus den Inhaltsbereich *Aufgaben und Material*

Abb. 2 Beispielitem aus dem Inhaltsbereich *Lernwege und Verständnis von Schüler:innen*

Welche Aussage über halbschriftliche Rechenstrategien trifft zu?

Kreuzen Sie bitte ein Kästchen an.

☐1 Halbschriftliche Strategien führen zu mehr Rechenfehlern.

☐2 Halbschriftliche Strategien unterstützen das flexible Rechnen.

☐3 Halbschriftlichen Strategien behindern das Kopfrechnen.

☐4 Halbschriftliche Strategien legen eine Notation fest.

☐0 Ich weiß die Antwort nicht.

den können, im Gegensatz zu den Päckchen in a) und c), bei denen unsystematisch Rechenoperationen eingeübt werden.

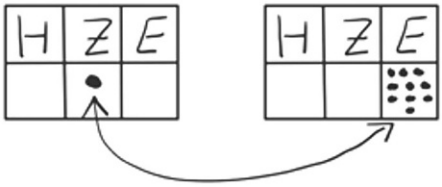
Abb. 2 zeigt ein Item aus dem Inhaltsbereich *Lernwege und Verständnis von Schüler:innen*. Hier wird mit der Relevanz halbschriftlicher Strategien ein wichtiger Aspekt des Mathematikunterrichts der Primarstufe adressiert. Dieses Thema lässt sich in der Leitidee *Zahl und Operation* verorten (Kultusministerkonferenz 2022).

Das in Abb. 3 dargestellte Item hat das Bündeln und Entbündeln im Dezimalsystem, welches ikonisch dargestellt wird, zum Inhalt. Der Itemstamm stellt einen typischen Inhalt des Mathematikunterrichts der Primarstufe dar.

Mit dem zu entwickelnden MaTe-Test (Assessment of *Ma*thematics Pedagogical Content Knowledge of Pre-Service Primary School *Te*achers) sollte es möglich sein, MPCK sowohl im Rahmen von größeren Studien, die eine IRT-Modellie-

Abb. 3 Beispielitem aus dem Inhaltsbereich *Lehren und Unterrichten*

Was wird in der Abbildung verdeutlicht?



Kreuzen Sie bitte ein Kästchen an.

☐1 Die Quersumme im Dezimalsystem

☐2 Die Subtraktion auf der Ebene von Stellenwerten

☐3 Das Bündeln und Entbündeln im Dezimalsystem

☐4 Die Begriffe Ziffernwert und Stellenwert

☐0 Ich weiß die Antwort nicht.

rung ermöglichen, als auch in Studien mit geringerer Stichprobengröße, die nach Klassischer Testtheorie Testwerte abbilden (Eid und Schmidt 2014), zu erfassen. Ebenso sollte eine sparsame Itemanzahl ermöglichen, dass der Test nicht im Multi-Matrix-Design eingesetzt werden muss, was eine Testwertmodellierung nach Klassischer Testtheorie nicht erlauben würde. Eine Übersicht der Iteminhalte findet sich in Anhang Tab. 5. Alle Testaufgaben wurden im Single Multiple Choice Format mit einem Attraktor und drei Distraktoren konstruiert, um das Gütekriterium der Ökonomie nicht nur hinsichtlich der Durchführung, sondern auch hinsichtlich der Auswertung zu erzielen. Jede Aufgabe enthielt zusätzlich die Option *Ich weiß die Antwort nicht*. Das Ankreuzen dieser Option oder eines Distraktors wurde beim Scoring mit 0 für *falsch* kodiert, wohingegen die Wahl des Attraktors im Scoring mit 1 für *richtig* kodiert wurde. Insgesamt wurden 110 Items konstruiert. Nach einer internen Reviewrunde im Projektteam wurden Items identifiziert, die den inhaltlichen und formalen Anforderungen der Testkonstruktion genügten. Ein Pool bestehend aus 59 Items wurde im Rahmen eines Expert:innenratings hinsichtlich der Güte der inhaltlichen Repräsentation auf Itemebene untersucht (Ausführungen hierzu siehe Abschn. 7.1). Items, denen nicht ausreichend inhaltliche Repräsentanz bescheinigt werden konnte, wurden eliminiert. Die übrigen 44 Items wurden in einer Pilotierung anhand von $n = 314$ Lehramtsstudierenden für die Primarstufe zweier Universitäten hinsichtlich ihrer empirischen Güte analysiert. Auf Basis der Pilotierungsergebnisse wurden 24 Items für die Testzusammenstellung selektiert. Vier Items beinhalteten Abbildungen, die für die Lösung der Testaufgabe berücksichtigt werden mussten (siehe beispielsweise das Item in Abb. 3). Eine Prä-Testung hinsichtlich der Bearbeitungsdauer der 24 Items ergab durchschnittlich $M = 10{,}06$ min ($SD = 1{,}68$), wobei mindestens acht Minuten und maximal 13,1 min Testzeit von den Proband:innen benötigt wurden.

5 Vorliegende Studie: Validierung des MaTe-Tests zum mathematikdidaktischen Wissen von Lehramtsstudierenden der Primarstufe

Übergeordnetes Ziel der vorliegenden Studie ist es, empirische Evidenzen im Rahmen einer umfassenden Validierung zu sammeln, die nach Kane (2001) eine Testwertinterpretation im Sinne der erklärenden Inferenz des MaTe-Tests erlaubt. Diese Testwertinterpretation wird dann gestützt, wenn die Testscores des MaTe-Tests als Indikator für das theoretische Konstrukt MPCK angenommen werden können. Entsprechend des argumentationsbasierten Ansatzes der Validierung nach Kane (2013) wurden folgende zu prüfende Grundannahmen formuliert:

Grundannahme 1: *Die Items des MaTe-Tests repräsentieren die für das mathematikdidaktische Wissen von Lehramtsstudierenden der Primarstufe relevanten Inhalte.*

Die Evidenz für diese Annahme kann in einer Inhaltsvalidierung erbracht werden, bei der Expert:innen die Güte der Repräsentanz des intendierten Iteminhalts einschätzen (Hartig et al. 2020; Jenßen et al. 2015).

Grundannahme 2: *Die MaTe-Testscores lassen sich auf eine einzige zugrunde liegende Fähigkeitsdimension zurückführen.*

Entsprechend der theoretischen Konzeption von MPCK und der Anforderung an den MaTe-Test, MPCK in einem Testwert abzubilden, kann diese Annahme in der Evidenzkategorie Struktur geprüft werden.

Grundannahme 3: *Die MaTe-Testscores zeigen differentielle Beziehungen zu anderen Variablen, die aus einem nomologischen Netz abgeleitet werden können.*

Unter Grundannahme 3 lassen sich mehrere Hypothesen formulieren, die allesamt innerhalb der Evidenzkategorie Beziehungen zu anderen Variablen geprüft werden können. Entsprechend der theoretischen Konzeptualisierung kann angenommen werden, dass Lehramtsstudierende für die Primarstufe ihr MPCK mit fortlaufendem Studium erwerben und dieser Erwerb mit Leistungen aus der Schulzeit konvergent bzw. divergent zusammenhängt. Ausgehend von theoretischen Annahmen (Ball et al. 2008; Baumert und Kunter 2006; Shulman 1986) lässt sich zudem ein konvergenter Zusammenhang zwischen MCK und MPCK annehmen.

Studien weisen darauf hin, dass diese beiden Wissensfacetten bei Lehrkräften eher hoch miteinander korrelieren. Die Studie von Voss et al. (2011) zeigte, dass MCK und MPCK von Referendar:innen für das Sekundarstufenlehramt latent hoch mit $r = 0{,}91$ korrelierte. Bezüglich des MCK und des MPCK von angehenden frühpädagogischen Fachkräften konnte eine latente Korrelation von $r = 0{,}73$ ermittelt werden (Blömeke et al. 2015b). Die TEDS-M-Studie von Blömeke et al. (2010a) speziell für Lehramtsstudierende der Primarstufe weist auf eine manifeste Korrelation von $r = 0{,}62$ zwischen MCK und MPCK hin. Die Höhe dieser Korrelation deutet darauf hin, dass sich beide Konstrukte nicht nur theoretisch, sondern auch empirisch voneinander unterscheiden. Dementsprechend lässt sich die folgende Hypothese formulieren:

a. Aufgrund der Konzeptualisierung des fachdidaktischen Wissens, die neben einer pädagogisch-psychologischen Perspektive eine fachlich geprägte Perspektive auf Mathematikunterricht der Primarstufe berücksichtigt, hängen die MaTe-Testscores hoch mit mathematischem Fachwissen von Lehramtsstudierenden der Primarstufe zusammen. Unter Berücksichtigung der Evidenzen für angehende Lehrkräfte in Deutschland und speziell dem Studienergebnis von Blömeke et al. (2010a) nehmen wir eine manifeste Korrelation an, die nicht über 0,9 liegt, sondern etwa bei 0,6.

Die folgenden Hypothesen beziehen sich auf die theoretische Annahme, dass kognitive Konstrukte bei Lernenden in der Regel miteinander zusammenhängen (Rindermann 2007), wobei domänenspezifische Variablen (z. B. Noten in Mathematik) einen Zusammenhang mit den MaTe-Testscores aufweisen und Variablen anderer Domänen (z. B. Noten in Deutsch) oder Variablen, die eine Fülle domänenspezifischer und domänenferner Leistungen umfassen (z. B. Abiturdurchschnittsnote) divergente Zusammenhänge zeigen sollten. Vor dem Hintergrund, dass höhere Noten eine geringere Leistung repräsentieren, können folgende Hypothesen abgeleitet werden:

b. Mit der letzten Schulnote in Mathematik zeigen sich mittlere negative Zusammenhänge. Diese Annahme lässt sich aufgrund der Domänenspezifität bezüglich der Inhalte formulieren. Da die Schulnote in Mathematik aber keinerlei didaktische Inhalte enthält, wird kein starker Zusammenhang angenommen.
c. Mit der letzten Schulnote in Deutsch zeigen sich keine bis kleine negative Zusammenhänge. Diese Annahme lässt sich dahingehend begründen, dass die letzte Schulnote in Deutsch als ein Proxy für Sprachkompetenz und Leseverständnis angesehen werden kann und Testaufgaben zum professionellen Wissen in der Regel längere Textteile beinhalten. Da der MaTe-Test keine Fallvignetten oder längere Situationsbeschreibungen beinhaltet, gehen wir davon aus, dass der Zusammenhang nur gering ausfällt.
d. Mit der Abiturdurchschnittsnote zeigen sich keine bis kleine negative Zusammenhänge. Studien weisen auf Zusammenhänge von Wissenstests zur Abiturdurchschnittsnote von Lehramtsstudierenden hin (z. B. Kleickmann et al. 2013), wobei diese auch dann gering ausfallen, wenn die Inhalte des Wissenstests den Inhalten des Abiturs entsprechen (z. B. Besser et al. 2021). Da der MaTe-Test Wissen erfassen soll, welches nicht während der Schulzeit erworben wird und die Abiturdurchschnittsnote neben einem Bündel an kognitiven Fähigkeiten aber auch affektiv-motivationale Dispositionen repräsentiert (Blömeke 2009), nehmen wir keinen Zusammenhang oder maximal einen kleinen Zusammenhang zwischen den beiden Variablen an.

Ausgehend von der theoretischen Konzeption professionellen Wissens lässt sich ableiten, dass mit zunehmendem Semester das professionelle Wissen ebenfalls zunimmt. Daher lässt sich folgende Hypothese formulieren:

e. Die MaTe-Testscores hängen positiv mit der Semesterzahl zusammen.

Zusammenfassend kann gesagt werden, dass die vorliegende Validierungsstudie anhand der Prüfung mehrerer Grundannahmen in den Kategorien *Inhalt, Struktur* und *Beziehungen zu anderen Variablen* untersucht, ob die MaTe-Testscores Schlussfolgerungen zum Konstrukt MPCK von Lehramtsstudierenden der Primarstufe im Sinne der Testwertinterpretation der Erklärung nach Kane (2001) erlauben.

6 Methode

6.1 Stichproben und Vorgehen

Für die Prüfung der *Grundannahme 1* in der Kategorie Inhalt wurden bereits während der Phase der Testentwicklung Professor:innen angefragt, die im Feld der mathematischen Bildung an Universitäten in Deutschland in Lehramtsstudiengängen lehren und die professionelle Kompetenz von Lehrkräften im Bereich Mathematik als einen ihrer Forschungsschwerpunkte haben. Insgesamt acht Professor:innen nahmen als Expert:innen an dem Rating teil. Die Expert:innen zeichneten sich durch ihre Berufstätigkeit im Hochschulbereich von $M = 16$ Jahren ($SD = 6$; $Min = 9$; $Max = 28$),

ihre allgemeine Publikationstätigkeit von $M = 35$ Publikationen in ihrer bisherigen Laufbahn ($SD = 28$; $Min = 13$; $Max = 100$) und ihrer Anzahl an Publikationen, die sich speziell auf MPCK beziehen ($M = 10$; $SD = 6$; $Min = 4$; $Max = 23$), aus. Alle Expert:innen gaben an, vertraut mit Testverfahren zur Erfassung professionellen Wissens zu sein.

Um *Grundannahme 2* zu prüfen, wurden $n = 440$ Lehramtsstudierende für die Primarstufe einer Universität mit dem MaTe-Test getestet. Die Mehrheit (80,1 %) der Teilnehmenden gab an, weiblichen Geschlechts zu sein (19,4 % männlich, 0,5 % divers). Die Teilnehmenden wurden aus allen Fachsemestern des Lehramtsstudiums rekrutiert, wobei sie sich mindestens im 3. Fachsemester befanden ($M = 4{,}92$, $SD = 1{,}64$). Dies spiegelt sich auch in der Streuung des Alters wider ($M = 24{,}40$, $SD = 5{,}90$). Die Teilnehmenden bearbeiteten online die 24 Items des MaTe-Tests, wofür diese mithilfe der Software *LimeSurvey* aufbereitet wurden. Die Studierenden erhielten keine Aufwandsentschädigung und nahmen freiwillig an der Erhebung teil.

Um die Hypothesen der *Grundannahme 3* überprüfen zu können, wurden Teilstichproben der vorangegangenen Gesamtstichprobe untersucht, da nicht alle Teilnehmenden der Gesamtstichprobe alle Instrumente bearbeitet haben. Teilstichprobe 1 umfasste $n_1 = 223$ Lehramtsstudierende für die Primarstufe und wurde für die Prüfung der Annahme 3a herangezogen. Die Teilnehmenden stammten alle aus einer Vorlesung zum Ende des Bachelorstudiums und hatten somit bereits an Lehrveranstaltungen zu fachlichen und fachdidaktischen Inhalten der Arithmetik, Geometrie und Stochastik teilgenommen. Teilstichprobe 2 bestand aus $n_2 = 217$ Lehramtsstudierenden der Primarstufe und wurde für die Prüfung der Annahmen 3b bis 3e genutzt. Die Mehrheit der Teilnehmenden befand sich im Bachelorstudium (85 %) und die Minderheit im Masterstudium (15 %). Hinsichtlich des Geschlechterverhältnisses und der Altersverteilung unterschieden sich die beiden Teilstichproben deskriptiv nicht von den Verteilungen der Gesamtstichprobe. Eine Übersicht der Stichproben, die zur Prüfung der Grundannahmen herangezogen wurden, liefert Tab. 1.

6.2 Instrumente

Alle teilnehmenden Lehramtsstudierenden bearbeiteten alle 24 Items des MaTe-Tests. Die Teilnehmenden der Teilstichprobe 1 bearbeiteten zusätzlich den KomMa-MCK-Test, um das fachmathematische Wissen zu erheben (Blömeke et al. 2015b). Dieser Test war ursprünglich zur Erfassung des fachmathematischen Wissens von frühpädagogischen Fachkräften vorgesehen, wurde jedoch inhaltlich auf dem Niveau des Primarstufenlehramts konstruiert (ebd.). Die Eignung des Tests für Lehramtsstudierende für die Primarstufe konnte entsprechend in einer früheren Studie

Tab. 1 Stichproben zur Prüfung der Grundannahmen

Grundannahme 1	Acht Expert:innen im Bereich der Mathematikdidaktik
Grundannahme 2	440 Lehramtsstudierende für die Primarstufe ab dem drittem Fachsemester
Grundannahme 3a	223 der 440 Studierenden aus einer Vorlesung im 5. Semester des Bachelorstudiums
Grundannahmen 3b–e	217 der 440 Studierenden, davon 85 % im Bachelor-Studium und 15 % im Master-Studium

gestützt werden (Dunekacke et al. 2014). Ebenso konnte gezeigt werden, dass der KomMa-MCK-Test Testwerte liefert, die sich von allgemein-kognitiven Fähigkeiten empirisch abgrenzen lassen (Jenßen et al. 2019). Der Test umfasst ursprünglich 24 Items im Single-Multiple-Choice-Format und im Open-Response-Format und erfasst das fachmathematische Wissen entlang vier verschiedener Inhaltsbereiche (Arithmetik, Geometrie, Stochastik sowie Größen, Muster und Strukturen) unter Berücksichtigung prozessbezogener Kompetenzen (Argumentieren, Darstellen, Modellieren, Problemlösen und Kommunizieren). Da die Erhebung online stattfand, wurden zwei Items des KomMa-MCK-Tests ausgeschlossen, da diese eine Zeichnung erforderten und dies bei der Aufbereitung via LimeSurvey nicht umgesetzt werden konnte. Alle Testaufgaben werden wie der MaTe-Test entlang des Richtig-Falsch-Schemas in 1 und 0 beim Scoring repräsentiert. Von den Teilnehmenden der Teilstichprobe 2 wurden neben den MaTe-Testscores und der Semesterzahl noch die letzte Schulnote in Mathematik, die letzte Schulnote in Deutsch sowie die Abiturdurchschnittsnote erfragt.

6.3 Analysen

Die Daten des Expert:inneratings wurden anhand der von Jenßen et al. (2015) vorgeschlagenen Prozedur zur Inhaltsvalidierung erhoben und analysiert. Die Expert:innen beantworteten auf einer vierstufigen Skala von 1 (= gar nicht) bis 4 (= voll und ganz) schriftlich die Frage, inwieweit der Iteminhalt den intendierten Inhalt und damit das Konstrukt MPCK für die Primarstufe optimal repräsentiert. Die Expert:innen konnten nicht erfahren, wie die anderen Expert:innen abgestimmt hatten. Zur Abfrage erhielten sie neben dem jeweiligen Item lediglich die Information, welcher der vier konzeptualisierten Inhaltsbereiche des mathematikdidaktischen Wissens für die Primarstufe das Item erfassen soll. Die Expert:innen konnten zu jedem ihrer Ratings Anmerkungen machen und ihr Rating begründen. Über alle Expert:innen wurden pro Item die Ratings gemittelt, damit alle Ratings gleichwertig in die Entscheidung über das Item eingehen konnten (Jenßen et al. 2015). Demnach müssen Items mit einem Mittelwert unter 2,5 eliminiert werden, wohingegen Items mit einem Mittelwert zwischen 2,5 und kleiner 3 anhand der Anmerkungen der Expert:innen überarbeitet werden können. Items mit einem Wert von mindestens 3 dürfen ohne Änderung für die Testzusammenstellung selektiert werden, da davon auszugehen ist, dass sie den intendierten Inhalt bereits hinreichend optimal repräsentieren. Das Vorgehen von Jenßen et al. (2015) berücksichtigt nicht die Streuung der Expert:innenratings. Eine Diskussion der Vor- und Nachteile dieses Vorgehens findet sich ebenfalls bei Jenßen et al. (2015).

Die Analysen zu den Grundannahmen 2 und 3 wurden IRT-basiert mithilfe der Software *ConQuest* durchgeführt (Wu et al. 1997). Grundannahme 2 wurde in der Kategorie Struktur untersucht, indem ein eindimensionales und ein vierdimensionales Raschmodell im Modellvergleich analysiert wurden. Zur Feststellung der Modellgültigkeit wurden die Informationskriterien AIC und BIC basierend auf Stichprobengröße, Parameteranzahl und Devianz berechnet (Pohl und Carstensen 2012), wobei geringere Werte die bessere Passung implizieren. Zur weiteren Feststellung der empirischen Güte des Modells wurden zur Evaluation die von Pohl und Cars-

tensen (2012) beschriebenen Fit-Kriterien herangezogen: Als erstes Maß wurden Weighted Mean Squares analysiert (WMNSQ: Wright und Masters 1982), wobei diese bei kleineren Stichproben nicht über 1,20 liegen sollten (konventionsgemäß hat sich durchgesetzt, dass der Wert nicht unter 0,80 liegen sollte; umso näher der Wert um 1,00 herum liegt, desto besser passt das angenommene Modell zu den Daten). Als weiteres Merkmal wurden die Diskriminationsparameter der Items herangezogen, wobei diese über 0,20 liegen sollten. Umso höher dieser Parameter liegt, desto besser repräsentiert das Item den Gesamttest. Zur Evaluation der Reliabilität wurden die Messgenauigkeit der Personenparameterschätzungen (WLE: Weighted Maximum Likelihood Estimates) und Cronbachs Alpha herangezogen, die von *ConQuest* ausgegeben werden.

Um die Grundannahme 3 zu untersuchen, wurden die Fähigkeitsparameter der Teilnehmenden, die aus der Skalierung im Rahmen der Prüfung von Grundannahme 2 gewonnen wurden, mit den Werten der jeweiligen Variablen korreliert. Die Interpretation der Stärke des Zusammenhangs folgte der Konvention nach Cohen (1988). Zur Analyse wurde die Software *Mplus 8* (Muthén und Muthén 2017) verwendet. Die Analyse mit *Mplus* erlaubt es, fehlende Werte mit der Full Information Maximum Likelihood (FIML) Methode zu berücksichtigen. Die fehlenden Werte hatten – je nach Zusammenhangsanalyse – einen Anteil von 0,5 bis 4,6 % (im Durchschnitt 1,8 %). Um zu prüfen, ob die fehlenden Werte zufällig fehlen, wurde der entsprechende Test von Little (1988) durchgeführt. Das Ergebnis wies darauf hin, dass die fehlenden Werte nicht systematisch auftraten ($\chi^2 = 32{,}25$, df = 23, $p = 0{,}095$).

7 Ergebnisse der Prüfungen der Annahmen

7.1 Prüfung Grundannahme 1

Da die Inhaltsvalidierung bereits im Prozess der Testentwicklung umgesetzt wurde und sich diese auf alle für die Pilotierung freigegebenen Items bezog, werden im Folgenden nur die Ergebnisse für die final selektierten Items berichtet. Die Mittelwerte der Ratings für die Items waren $Min = 2{,}5$ und $Max = 4$ mit einem mittleren Rating über alle Items von $M = 3{,}35$ ($SD = 0{,}44$). Vier Items befanden sich in der nach Jenßen et al. (2015) beschriebenen Kategorie der Items, die überarbeitet werden müssen, da sie den Inhalt nicht optimal repräsentierten. Die übrigen 20 Items befanden sich in der obersten Kategorie, in der die Items bereits ohne Überarbeitung angenommen werden können. Zur Überarbeitung der Items wurden die Anmerkungen der Expert:innen herangezogen. Diese bezogen sich entweder auf konkrete Formulierungen oder die mangelnde Repräsentation des Items in Bezug auf einen der vier Inhaltsbereiche, nicht jedoch auf einen Mangel an Repräsentation für das Konstrukt MPCK für die Primarstufe.

7.2 Prüfung Grundannahme 2

Das vierdimensionale Modell, in dem jede latente Dimension einen inhaltlichen Bereich widerspiegelte, zeigte latente Korrelationen zwischen den Dimensionen im

Tab. 2 Informationskriterien für Modellvergleich

Modell	Anzahl der Parameter	Devianz	AIC	BIC
1-dimensionales Modell	25	9832	9882	9984
4-dimensionales Modell	34	9821	9889	10027

Tab. 3 Deskriptive Ergebnisse

	M	*SD*
KomMa-MCK-Test	15,93	3,70
Letzte Schulnote Mathematik	2,46	1,02
Letzte Schulnote Deutsch	2,33	0,83
Abiturdurchschnittsnote	2,46	1,96

Bereich von $r = 0{,}90$ und $r = 0{,}96$. Die WLE-Reliabilitäten lagen zwischen 0,02 und 0,44. Entsprechend zeigte sich auch bei der Analyse der Informationskriterien (siehe Tab. 2) die bessere Passung des eindimensionalen Modells. Die Annahme eines vierdimensionalen Modells wurde damit verworfen.

Die Reliabilitäten im eindimensionalen Modell lagen bei WLE-Rel. = 0,72 und Cronbachs Alpha = 0,74. Die Varianz der latenten Dimension lag bei $Var = 0{,}83$. Die WMNSQ-Werte lagen zwischen 0,88 und 1,07 ($M = 1{,}00$, $SD = 0{,}05$) und deuteten ebenso wie die Diskriminationsparameter mit $M = 0{,}38$ ($SD = 0{,}08$), $Min = 0{,}20$, $Max = 0{,}55$ auf einen guten Fit zwischen den MaTe-Testscores und dem eindimensionalen Raschmodell hin.

Die latente Verteilung der Itemparameter ist in Anhang B dargestellt. Die Itemmittelwerte lagen zwischen $Min = 0{,}25$ ($SD = 0{,}43$) und $Max = 0{,}95$ ($SD = 0{,}22$). Der Mittelwert über alle Itemmittelwerte lag bei $M = 0{,}67$ ($SD = 0{,}23$) mit einem Median von $Md = 0{,}72$.

7.3 Prüfung Grundannahme 3

In Tab. 3 sind die deskriptiven Ergebnisse für die weiteren Variablen angegeben. In Tab. 4 sind die Korrelationen dieser Variablen mit dem MaTe-Testscore dargestellt. Es zeigt sich, dass Annahme 3a (hoher positiver Zusammenhang mit MCK) und 3d (kleiner oder kein Zusammenhang mit der Abiturdurchschnittsnote) als bestätigt anzusehen sind. Ebenso konnten Annahme 3b (mittlerer Zusammenhang mit der letzten Schulnote in Mathematik) sowie Annahme 3e (positiver Zusammenhang mit der Semesterzahl) bestätigt werden. Für die Annahme 3c (kleiner Zusammenhang mit der letzten Schulnote in Deutsch) zeigte sich jedoch keine bestätigende Evidenz.

Tab. 4 Korrelationen der Variablen mit dem MaTe-Testscore

	KomMa-MCK-Test	Letzte Schulnote Mathematik	Letzte Schulnote Deutsch	Abiturdurch-schnittsnote	Semester
MaTe-Testscore	0,58 ($p < 0{,}01$)	–0,26 ($p < 0{,}01$)	–0,08 ($p = 0{,}09$)	–0,05 ($p = 0{,}46$)	0,36 ($p < 0{,}01$)

8 Zusammenfassung und Diskussion

Ausgehend von den dargestellten Analysen der vorliegenden Validierungsstudie zeigte sich, dass der MaTe-Test einen globalen Testwert liefert, der eine reliable Schätzung von WLEs ermöglicht. Die in der vorliegenden Studie ermittelte Reliabilität von Cronbachs Alpha = 0,74 mag vor dem Hintergrund der Natur des Konstrukts (MPCK wird üblicherweise als ein eher heterogenes Merkmal im Vergleich zu anderen Leistungskonstrukten verstanden wie sie in der psychologischen Diagnostik oft Gegenstand sind) und dem intendierten Testeinsatz (MaTe als ökonomischer Test, der nicht zu Zwecken der Individualdiagnostik entwickelt wurde) als hinreichend angesehen werden (Gäde et al. 2020). Für Anwendungen im Längsschnitt mag eine höhere Reliabilität nötig sein. Die Ökonomie des Tests zeigt sich nicht nur bezüglich der Itemanzahl und dem Antwortformat (24 Single-Multiple-Choice-Items) sondern auch in einer daraus resultierenden geringen Bearbeitungszeit von etwa 10 min.

Itemmittelwerte sowie Median deuten darauf hin, dass der MaTe-Test hinsichtlich seiner Schwierigkeit eher als leicht angesehen werden kann. Eine Differenzierung im oberen Bereich der Ausprägung ist damit auch in zukünftigen Studien eher eingeschränkt, was beim spezifischen Testeinsatz berücksichtigt werden sollte.

Wie von Messick (1989) und der AERA, APA und NCME (2014) empfohlen, wird im Folgenden eine integrative Betrachtung der vorliegenden Validitätshinweise vorgenommen. Ziel war es, zu untersuchen, ob die Testwerte des MaTe-Tests im Sinne der *Erklärung* interpretiert werden können. *Erklärung* meint hier, dass die Testwerte das Konstrukt MPCK von Lehramtsstudierenden für die Primarstufe widerspiegeln. Der Test ist dabei eher als begleitendes Instrument bzw. der Testwert als Kovariate in weiteren Studien geeignet. Beispielsweise kann der MaTe-Test eingesetzt werden, um Entwicklungen anderer Variablen unter Kontrolle des fachdidaktischen Wissens zu untersuchen oder um bei qualitativen Studien mit kleinen Stichproben (z. B. auch im Rahmen von Interventionen), in denen situationsspezifische Fertigkeiten bzw. die Performanz von (angehenden) Lehrkräften untersucht werden, einen Proxy für das fachdidaktische Wissen der Teilnehmenden zu erfassen.

Die angestrebte Testwertinterpretation wurde anhand von drei a priori formulierten Grundannahmen geprüft. Zur Prüfung der Grundannahme 1 in der Kategorie *Inhalt* wurde mit dem Itemrating ein gängiges Verfahren der Inhaltsvalidierung gewählt, wobei die Qualität des Verfahrens durch die von Jenßen et al. (2015) beschriebenen Kriterien zur Feststellung der Expertise der Rater:innen als hoch angesehen werden kann. Die Analyse zeigte, dass einzelne Items im Prozess der Itementwicklung noch nicht in optimaler Weise den intendierten Inhalt repräsentierten. Diese Items konnten mithilfe der konkreten Anmerkungen der Expert:innen überarbeitet werden. Trotz dieser vier Fälle mit ungenügender Passung deutete das gemittelte Rating über alle Items hinweg auf eine gute inhaltliche Repräsentanz hin. Allerdings wurde im Rahmen des Expert:innenratings nicht explizit nach Konstruktunter- oder -überrepräsentanz gefragt (Hartig et al. 2020). Der MaTe-Test enthält zwei Items zur Didaktik bei der Bruchrechnung, die jedoch in 14 von 16 Bundesländern nicht Teil des Rahmenlehrplans sind, wenn auch im Studium für Lehramtsstudierende der Primarstufe durchaus adressiert werden können. Somit könnte je nach Perspektive von einer Konstruktüber- oder unterrepräsentanz ausgegangen werden. Hinsichtlich

der Konstruktrepräsentanz kann aber auch festgestellt werden, dass zu allen als Heuristik für die Itementwicklung vorab formulierten Inhaltsbereichen Testaufgaben im MaTe-Test enthalten sind.

Hinsichtlich der Grundannahme 2 in der Evidenzkategorie *Struktur* zeigte ein eindimensionales Modell die beste Passung, so wie es ursprünglich angenommen wurde. Somit liefert der MaTe-Test einen globalen und reliablen Testwert, der die Antworten auf eine Fähigkeitsdimension zurückführt.

Hinsichtlich der *Beziehungen zu anderen Variablen* vor dem Hintergrund der Grundannahme 3 zeigen sich überwiegend hypothesenkonforme Eigenschaften des Testwerts. Das Korrelationsmuster entsprach den vorab formulierten Hypothesen. Im Falle der letzten Schulnote in Deutsch deuten unsere Daten darauf hin, dass die Bearbeitung der MaTe-Testaufgaben nicht mit sprachlichen Leistungen seitens der Teilnehmenden assoziiert sind. Hier würde sich zukünftig eine Validierung mit einem Testverfahren zur Messung sprachlicher Kompetenzen anbieten, um den Befund abzusichern. Ob sich bei dem nicht-signifikanten Ergebnis hinsichtlich der Abiturdurchschnittsnote – verstanden als Produkte von Lernprozessen über alle Fächer hinweg (Prenzel et al. 2007) – bereits zeigt, dass der MaTe-Test Aufgaben enthält, deren Inhalte nicht in der Schule und ausschließlich im Lehramtsstudium erworben werden, ist fraglich. Unter Hinzunahme des gefundenen Zusammenhangs mit der Semesterzahl wird diese Annahme gestützt, wenn auch das Design keine Falsifikation dieser Annahme erlaubt (Hartig et al. 2020). Hier bräuchte es zukünftig Längsschnittstudien, um dieses Validitätsargument zu stärken.

Alle drei Grundannahmen konnten in der vorliegenden Studie geprüft und bestätigt werden. Daher kann bisher davon ausgegangen werden, dass die MaTe-Testitems Aufgaben darstellen, die das Konstrukt MPCK von Lehramtsstudierenden der Primarstufe repräsentieren. Wir möchten an dieser Stelle explizit darauf verweisen, dass wir lediglich Grundannahmen zur Testwertinterpretation der Erklärung geprüft haben. Die Gültigkeit anderer Testwertinterpretationen wie Bewertung oder Extrapolation wurde nicht untersucht und der MaTe-Test kann daher nicht zuverlässig außerhalb des hier untersuchten Anwendungsfalls eingesetzt werden, sondern erst nach Prüfung entsprechender Grundannahmen (Kane 2001).

Die Ergebnisse der vorliegenden Validierungsstudie sind vor dem Hintergrund einiger Limitationen zu sehen. So wurden in den Studien nur Studierende einer Universität untersucht, was die Generalisierbarkeit auf andere Hochschulkontexte, z. B. in anderen Bundesländern mit abweichenden Curricula, beschränkt. Zudem lagen eher geringe Stichprobengrößen vor, was bei IRT-Modellierungen zu Verzerrungen in den Parameterschätzungen führen kann (Rost 2004). Als ebenfalls einschränkend kann unsere Variablenauswahl betrachtet werden, was wir am Beispiel der Erlernbarkeit von MPCK skizzieren möchten: Wir haben in der Studie den Zusammenhang mit der Semesterzahl untersucht, ein passenderes Vorgehen wäre jedoch die Analyse der Entwicklung des MPCK, welche explizit durch spezifische Lerngelegenheiten hervorgerufen werden müsste. Lerngelegenheiten wurden jedoch nicht im Rahmen der vorliegenden Validierung erhoben. Zukünftige Studien sollten diesen Ansatz jedoch einbeziehen, um Effekte von Lerngelegenheiten zu untersuchen.

Einschränkend lässt sich zudem festhalten, dass der MaTe-Test für die Erforschung der Entwicklung von MPCK von Lehramtsstudierenden für die Primarstufe

aus einer eher dynamische Perspektive heraus (Depaepe et al. 2013) weniger geeignet sein könnte, da er nur einen globalen Testwert bietet. Kognitive Anforderungen, die in pädagogische Situationen bestehen und sich entsprechend in Testformaten wiederfinden sollten (Krathwohl 2002), würden aufgrund des einheitlichen Single-Multiple-Choice-Formats zu gering variieren. Hier wären situierte Textvignetten eine mögliche Herangehensweise, um stärker das Handlungswissen von angehenden Lehrkräften zu erfassen. Dies hätte aber wiederum Auswirkungen auf die Ökonomie und Reliabilität des Tests, sodass dieser Ansatz nicht gewählt wurde.

Vorteilhaft an der vorliegenden Validierungsstudie ist, dass mehrere Evidenzkategorien entsprechend der Teststandards nach AERA, APA und NCME (2014) untersucht wurden, nämlich *Inhalt, Struktur* und mehrere *Beziehungen zu anderen Variablen*. Dies unterstreicht bereits die Güte des Validierungsprozesses (Hogan und Agnello 2004; Cizek et al. 2008). Dennoch sei darauf verwiesen, dass der Validierungsprozess nie abgeschlossen ist (Kane 2013; AERA, APA und NCME 2014), da immer wieder Validierungen anderer Schlussfolgerungen möglich sind, die bisher nicht untersucht wurden. Aus unserer Perspektive stellen die folgenden zukünftigen Validierungsstrategien sinnvolle Erweiterungen der vorliegenden Validierungsstudie dar: Auf Itemebene kann es interessant sein, Antwortprozesse in Form von Cognitive Labs zu untersuchen, um zu analysieren, welches spezifische mathematikdidaktische Wissen zur Lösung herangezogen wird (Padilla und Leighton 2017). Ebenso böte sich die Untersuchung der differentiellen Validität an (Kubinger 2019). Hier steht die Frage im Vordergrund, ob andere Berufsgruppen wie z. B. Erzieher:innen oder Sekundarstufenlehrkräfte andere Ergebnisse beim MaTe-Test erzielen. Vergleichbar zur COACTIV Studie könnte beispielweise vor dem Hintergrund der dort formulierten *Professional Knowledge Hypothesis* (Krauss et al. 2008a) untersucht werden, ob Erwachsene der Allgemeinbevölkerung schlechter beim MaTe-Test abschneiden als Lehramtsstudierende für die Primarstufe, da es sich um Wissen handelt, welches speziell im Studium erworben wird. Ebenso könnte untersucht werden, wie Ergebnisse des MaTe-Tests speziell nach Praxisphasen ausfallen, da es sich hier um spezielle Lerngelegenheiten handelt, denen im Rahmen des Lehramtsstudiums große Effekte bezüglich des Wissenserwerbs beigemessen werden (Phelps et al. 2019).

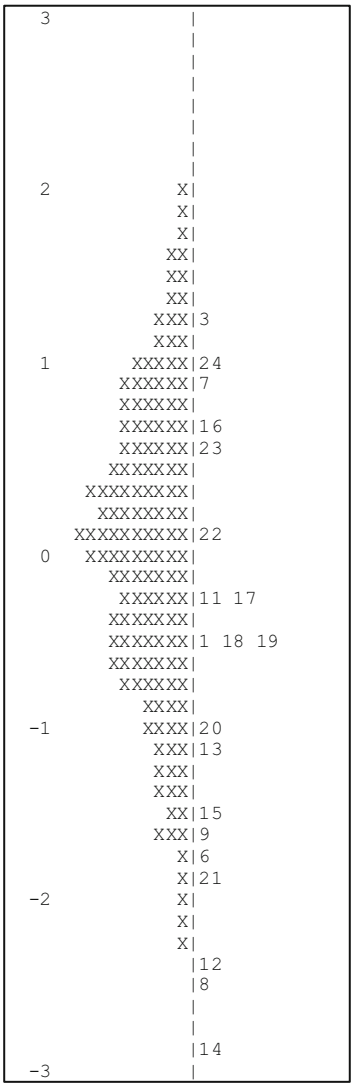
Im Sinne des wissenschaftspraktischen Ziels der Studie zur Schließung des Desiderats einen frei verfügbaren Test zu konstruieren, welcher das mathematikdidaktische Wissen von Lehramtsstudierenden für die Primarstufe ökonomisch, reliabel und valide erfasst, wird der vorliegende MaTe-Test der *Scientific Community* auf Anfrage bei den Autor:innen für weitere Forschungsprojekte zur Verfügung gestellt.

9 Anhang

Tab. 5 Testinhalte

Subdimension (Itemanzahl)	Iteminhalte
Lernwege und Verständnis von Schüler:innen (10)	Fehlertyp beim schriftlichen Addieren
	Grundvorstellungen zu Bruchzahlen
	Zahlaspekte
	Didaktische Prinzipien bei Rechenschwierigkeiten
	Halbschriftliche Rechenstrategien
	Fehler bei der Addition von Brüchen
	Grundvorstellungen zur Subtraktion
	Bruchzahlaspekt
	Halbschriftliches Addieren
	Grundvorstellungen zur Multiplikation
Aufgaben und Material (6)	Rechenstrategien beim Addieren
	Halbschriftliches Rechnen
	Aufgabenpäckchen zur Leitidee „Muster und Strukturen“
	Kapitänsaufgaben
	Digitale Medien im Mathematikunterricht
	Vermeidung von Ergebnisorientierung
Curricula und Lernkontexte (5)	Leitidee Raum und Form
	Leitidee Größen und Messen
	Allgemeine mathematische Kompetenzen in den Bildungsstandards für den Primarbereich
	Allgemeine mathematische Kompetenzen in den Bildungsstandards für den Primarbereich
	Zahlaspekte
Lehren und Unterrichten (3)	Stellenwerttafel
	EIS-Prinzip
	Zahlaspekte

Abb. 4 Latente Verteilung der Itemparameter



Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. American Psychological Association. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education

Agathangelou, S. A., & Charalambous, C. Y. (2020). Is content knowledge pre-requisite of pedagogical content knowledge? An empirical investigation. *Journal of Mathematics Teacher Education*. https://doi.org/10.1007/s10857-020-09466-0.

Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407. https://doi.org/10.1177/0022487108324554.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift Für Erziehungswissenschaft*, *9*(4), 469–520. https://doi.org/10.1007/s11618-006-0165-2.

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*(1), 133–180. https://doi.org/10.3102/0002831209345157.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research*, *35*(5), 463–482.

Besser, M., Göller, R., Ehmke, T., Leiss, D., & Hagena, M. (2021). Entwicklung eines fachspezifischen Kenntnistests zur Erfassung mathematischen Vorwissens von Bewerberinnen und Bewerbern auf ein Mathematik-Lehramtsstudium. *Journal für Mathematik-Didaktik*, *42*, 335–365.

Blömeke, S. (2009). Ausbildungs- und Berufserfolg im Lehramtsstudium im Vergleich zum Diplom-Studium – Zur prognostischen Valididtät kognitiver und psycho-motivationaler Auswahlkriterien. *Zeitschrift für Erziehungswissenschaft*, *12*, 82–110.

Blömeke, S., Kaiser, G., Döhrmann, M., Suhl, U., & Lehmann, R. (2010a). Mathematisches und mathematikdidaktisches Wissen angehender Primarstufenlehrkräfte im internationalen Vergleich. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich* (S. 197–251). Waxmann.

Blömeke, S., Kaiser, G., & Lehmann, R. (2010b). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich*. Waxmann.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015a). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift Für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194.

Blömeke, S., Jenßen, L., Dunekacke, S., Suhl, U., Grassmann, M., & Wedekind, H. (2015b). Leistungstests zur Messung der Professionellen Kompetenz Frühpädagogischer Fachkräfte. *Zeitschrift Für Pädagogische Psychologie*, *29*(3–4), 177–191. https://doi.org/10.1024/1010-0652/a000159.

Blömeke, S., Jenßen, L., & Eid, M. (2022). The role of intelligence and self-concept for teachers' competence. *Journal of Intelligence*, *10*(2), 20. https://doi.org/10.3390/jintelligence10020020.

Buchholtz, N., Kaiser, G., & Blömeke, S. (2014). Die Erhebung mathematikdidaktischen Wissens – Konzeptualisierung einer komplexen Domäne. *Journal für Mathematikdidaktik*, *35*(1), 101–128. https://doi.org/10.1007/s13138-013-0057-y.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.-P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: an experimental study. *American Educational Research Journal*, *26*(4), 499–531. https://doi.org/10.3102/00028312026004499.

Chen, R.-J. (2017). Prospective elementary teachers' aestetic experience and relationships to mathematics. *Journal of Mathematics Teacher Education*, *20*, 207–230. https://doi.org/10.1007/s10857-015-9329-4.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*(3), 397–412. https://doi.org/10.1177/0013164407310130.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Erlbaum.

Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education*, *34*, 12–25.

Dunekacke, S., Jenßen, L., Eilerts, K., & Grassmann, M. (2014). *Operationalisierung des mathematikbezogenen Wissens angehender frühpädagogischer Fachkräfte* (S. 1–15).
Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Hogrefe.
Gäde, J. C., Schermelleh-Engel, K., & Werner, C. S. (2020). Klassische Methoden der Reliabilitätsschätzung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 305–334). Springer.
Graeber, A., & Tirosh, D. (2008). Pedagogical content knowledge: Useful concept or elusive notion. In P. Sullivan & T. Wood (Hrsg.), *Knowledge and beliefs in mathematics teaching and teaching development. The international handbook of mathematics teacher education* (S. 117–132). Sense.
Hartig, J., Frey, A., & Jude, N. (2020). Validität von Testwertinterpretationen. In H. Mossbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 529–545). Springer Nature.
Hill, H. C., Loewenberg Ball, D., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conzeptualizing and measuring teachers' topics-specific knowledge of students. *Journal for Resarch in Mathematics Education*, *39*(4), 372–400.
Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, *64*(5), 802–812. https://doi.org/10.1177/0013164404264120.
Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Kompetenzen von Studierenden. In *Qualitätssicherung in der Kompetenzforschung*. Beiheft Der Zeitschrift für Pädagogik, (Bd. 61, S. 11–31).
Jenßen, L., Dunekacke, S., Gustafsson, J.-E., & Blömeke, S. (2019). Intelligence and knowledge: the relationship between preschool teachers' cognitive dispositions in the field of mathematics. *Zeitschrift für Erziehungswissenschaft*, *22*, 1313–1332. https://doi.org/10.1007/s11618-019-00911-2.
Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12001.
Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., & Baumert, J. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education. *Journal of Teacher Education*, *64*(1), 90–106.
Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Journal of Psychology*, *216*(2), 61–73. https://doi.org/10.1027/0044-3409.216.2.61.
König, J., Blömeke, S., & Kaiser, G. (2010). Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich* (S. 99–130). Waxmann.
Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory and Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2.
Krauss, S., Baumert, J., & Blum, W. (2008a). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *ZDM – International Journal on Mathematics Education*, *40*(5), 873–892. https://doi.org/10.1007/s11858-008-0141-9.
Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008b). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*(3), 716–725. https://doi.org/10.1037/0022-0663.100.3.716.
Kubinger, K. D. (2019). *Psychologische Diagnostik*. Hogrefe.
Kultusministerkonferenz (Hrsg.). (2022). Bildungsstandards für das Fach Mathematik Primarbereich. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2022/2022_06_23-Bista-Primarbereich-Mathe.pdf. Zugegriffen: 21.06.2023.
Liljedahl, P., Durand-Guerrier, V., Winslow, C., Bloch, I., Huckstep, P., & Rowland, T., et al. (2009). Components of mathematics teacher training. In R. Even & D. Loewenberg Ball (Hrsg.), *The professional education and development of teachers of mathematics. The 15th ICMI Study* (S. 25–34). Springer.
Little, R. J. A. (1988). A test of missing completeley at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202.
Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (S. 13–104). American Council on Education, Macmillan.
Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (pp. 1487–1495). MacMillan.
Muthén, L., & Muthén, B. (2017). *Mplus user's guide* (8. Aufl.). Muthén & Muthén. https://doi.org/10.1111/j.1600-0447.2011.01711.x.

Neumann, K., Kind, V., & Harms, U. (2019). Probing the amalgam: the relationship between science teachers' content, pedagogical and pedagogical content knowledge. *International Journal of Science Education*, *41*(7), 847–861. https://doi.org/10.1080/09500693.2018.1497217.

Padilla, J.-L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo & A. M. Hubley (Hrsg.), *Understanding and investigating response processes in validation research* (S. 211–228). Springer. https://doi.org/10.1007/978-3-319-56129-5.

Phelps, G., Howell, H., & Liu, S. (2019). Exploring differences in mathematical knowledge for teaching for prospective and practicing teachers. *ZDM—Mathematics Education*. https://doi.org/10.1007/s11858-019-01097-x.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report. Scaling the data of the competence tests*. NEPS working paper, Bd. 14. Otto-Friedrich-Universität, Nationales Bildungspanel.

Prenzel, M., Walter, O., & Frey, A. (2007). PISA misst Kompetenzen. Eine Replik auf Rindermann (2006). Was messen internationale Schulleistungsstudien? *Psychologische Rundschau*, *58*(2), 128–136. https://doi.org/10.1026/0033-3042.58.2.128.

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, *21*, 667–706. https://doi.org/10.1002/per.634.

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

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.30827/profesorado.v23i3.11230.

Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Review*, *57*, 1–21.

Thorsen, C., Gustafsson, J.-E., & Cliffordson, C. (2014). The influence of fluid and crystallized intelligence on the development of knowledge and skills. *British Journal of Educational Psychology*, *84*, 556–570. https://doi.org/10.1111/bjep.12041.

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology*, *103*(4), 952–969.

Weinert, F. E. (2001). Concept of competence: A conceptual classification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies*. Hogrefe.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ACER Conquest: Generalised item response modelling software*. ACER.