Originalarbeit

Der BilWiss-2.0-Test

Ein revidierter Test zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften

Olga Kunina-Habenicht[1], Christina Maurer[2], Kristin Wolf[2], Doris Holzberger[3], Maria Schmidt[3], Theresa Dicke[4], Ziwen Teuber[5], Marta Koc-Januchta[6], Hendrik Lohse-Bossenz[7], Detlev Leutner[6], Tina Seidel[3] und Mareike Kunter[2]

[1]Institut für bildungswissenschaftliche Forschungsmethoden, Pädagogische Hochschule Karlsruhe
[2]Institut für Psychologie, Goethe-Universität Frankfurt am Main
[3]TUM School of Education, Technische Universität München
[4]Institute for Positive Psychology and Education, Australian Catholic University
[5]Fakultät für Psychologie und Sportwissenschaft, Universität Bielefeld
[6]Fakultät für Bildungswissenschaften, Universität Duisburg-Essen
[7]Pädagogische Hochschule Heidelberg

Zusammenfassung: In diesem Beitrag wird ein revidierter Test zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften, der BilWiss-2.0-Test, vorgestellt, und es werden Hinweise auf die psychometrische Güte der mit dem Instrument gemessenen Testwerte präsentiert. Das bildungswissenschaftliche Wissen umfasst neben unterrichtsnahen Inhalten auch Kenntnisse bspw. über Bildungstheorie, Schulorganisation oder Wissen über den Lehrerberuf. Die Kurzform des revidierten Wissenstests beinhaltet 65 Items, die Langform des Tests 119 Items aus sechs verschiedenen Inhaltsbereichen. Auf der Datengrundlage von 788 Lehramtsstudierenden mehrerer Universitäten wurden 2-PL-Partial-Credit-IRT-Modelle geschätzt. Bezüglich der faktoriellen Struktur zeigte sich in Strukturgleichungsmodellen für die Kurzform, dass sich die sechs Inhaltsbereiche gut als sechs untereinander korrelierende latente Faktoren abbilden lassen. Für die konvergente und prognostische Validität der Testwertinterpretationen sprechen a) schwache statistisch signifikante Korrelationen zwischen der Testleistung und der Anzahl der besuchten relevanten inhaltsspezifischen Lehrveranstaltungen und Studienleistungen und b) statistisch signifikant schlechtere Wissensleistungen von Studierenden, die mindestens einmal eine Prüfung wiederholen mussten, im Vergleich zu Personen, die keinen Wiederholungsversuch benötigten.

Schlüsselwörter: Lehramtsausbildung, Professionswissen, Wissenstest, bildungswissenschaftliches Wissen, Testentwicklung, Validität

The BilWiss-2.0 Test: A Revised Instrument for the Assessment of Teachers' Educational Knowledge

Abstract: We describe the development of the revised BilWiss-2.0 Test, assessing generic educational knowledge and we present evidence for the reliability of the test and the validity of the test score interpretations. Educational knowledge covers knowledge domains that are directly related to instruction and other potentially relevant distant knowledge aspects such as educational theory and knowledge about the teaching profession. The short test version includes 65 items, whereas the long version contains 119 items from six knowledge domains. We applied 2-PL partial-credit item response theory (IRT) models to a data set collected from 788 teacher students from different German universities. With regard to the empirical structure of the test, structural equation models indicated a good fit of the model with six correlated latent factors. Small significant correlations between test performance and the number of relevant university courses attended and grades attained during university studies support the convergent and prognostic validity of test score interpretations. Moreover, students who repeated at least one examination showed significantly lower test performance than students who passed on the first try.

Keywords: Teacher education, professional knowledge, knowledge test, educational knowledge, test development, validity

Aktuelle Ansätze zur Professionalität im Lehrerberuf gehen davon aus, dass die Fähigkeit, qualitativ hochwertige Bildungsarbeit zu leisten, das Ergebnis eines professionellen Entwicklungsprozesses ist (Baumert & Kunter, 2006; Schleicher, 2016). Dabei besteht Konsens, dass das professionsspezifische Wissen von Lehrkräften einen wichtigen Aspekt ihrer professionellen Kompetenz darstellt und in einem berufsbiografischen Prozess aufgebaut werden kann (Baumert et al., 2010; Shulman, 1986). Die vielfältigen Anforderungen an den Lehrerberuf spiegeln sich in

der Heterogenität dieses professionellen Wissens wider: Traditionell wird zwischen Fachwissen, fachdidaktischem Wissen und pädagogischem Wissen unterschieden (Baumert & Kunter, 2006; Shulman, 1986; Voss, Kunina-Habenicht, Hoehne & Kunter, 2015).

Der Aufbau einer solchen Wissensbasis soll schon bei Lehramtsstudierenden während der ersten, universitären Phase der Lehrerbildung beginnen. Dementsprechend gliedert sich das Studium zum einen in einen fachwissenschaftlichen und fachdidaktischen Anteil, zum anderen in die fachunabhängigen, generisch orientierten Bildungswissenschaften, an denen in der Regel Disziplinen wie Psychologie, Erziehungswissenschaft und Soziologie beteiligt sind (Terhart, 2012). Die Curricula des bildungswissenschaftlichen Studienteils orientieren sich dabei zunehmend an den von der Kultusministerkonferenz (TKMK, 2014) verabschiedeten Standards für die Lehrerbildung in den Bildungswissenschaften (Unterrichten, Erziehen, Beurteilen und Innovieren; Hohenstein, Zimmermann, Kleickmann, Köller & Möller, 2014). Unter bildungswissenschaftlichem Wissen verstehen wir dementsprechend das fachunabhängige Professionswissen von Lehrkräften als wichtige Voraussetzung für das professionelle Agieren im Schulkontext innerhalb und außerhalb des Unterrichts (vgl. Kunter et al., 2017; Linninger et al., 2015).

Ein zentrales Anliegen der empirischen Forschung zur professionellen Kompetenz von Lehrkräften ist es, praxisrelevante Erträge des Professionswissens sichtbar zu machen (Baumert & Kunter, 2006). Zur objektiven Erfassung des Fachwissens und fachdidaktischen Wissens wurden standardisierte Testverfahren für verschiedene Fächer entwickelt (bspw. Baumert et al., 2010; Blömeke, Bremerich-Vos, Kaiser, Nold & Schwippert, 2013; Jüttner, Boone, Park & Neuhaus, 2013; Kleickmann et al., 2014). Bisherige empirische Ergebnisse weisen darauf hin, dass fachdidaktisches Wissen die Unterrichtsqualität und indirekt die Leistungen von Schülerinnen und Schülern positiv beeinflusst (Baumert et al., 2010; Kunter et al., 2013; Lange et al., 2015).

Für das fachunabhängige bildungswissenschaftliche Wissen ist die Evidenzlage dagegen deutlich dünner, obwohl die Prüfung der Praxisrelevanz des entsprechenden Studienteils – gerade angesichts der Kritik in Bezug auf seine vermeintliche Praxisferne und inhaltliche Beliebigkeit bzw. mangelhafte Kumulativität – ein wichtiges Forschungsdesiderat darstellt (Alles, Apel, Seidel & Stürmer, 2018; Darling-Hammond, Chung & Frelow, 2002; Veenman, 1984). In aktuellen Forschungsvorhaben wurden daher in den letzten Jahren verschiedene standardisierte Testinstrumente zur Untersuchung des bildungswissenschaftlichen Wissens von Lehrkräften entwickelt (für eine Übersicht siehe Voss et al., 2015). Ziel dieses Beitrags ist die Vorstellung eines revidierten Tests, welcher das bildungswissenschaftliche Wissen in seiner inhaltlichen Breite ökonomisch erfasst. Dargestellt werden die Entwicklung und Inhaltsbereiche des Tests sowie empirische Ergebnisse zur Teststruktur und erste Hinweise auf die Validität der Testwertinterpretationen.

Erfassung des bildungswissenschaftlichen Wissens von Lehrkräften

In den letzten zehn Jahren wurden mehrere Verfahren zur Erfassung des unterrichtsbezogenen, pädagogisch-psychologischen Wissens von Lehrkräften entwickelt (u. a. König & Blömeke, 2009; Lenske, Thillmann, Wirth, Dicke & Leutner, 2015; Seifert, Hilligus & Schaper, 2009; Voss, Kunter & Baumert, 2011). Die Tests sprechen zum einen deklaratives Faktenwissen, zum anderen prozeduralisiertes Handlungswissen an (für eine Übersicht siehe Voss et al., 2015). Der Test zum pädagogisch-psychologischen Wissen von Voss et al. (2011) bspw. fokussiert mit den Dimensionen *Klassenführung, Unterrichtsmethoden, Leistungsbeurteilung* und *Individuelle Lernprozesse* auf unmittelbar unterrichtsrelevante Wissensbestände, die primär pädagogisch-psychologisch fundiert sind. Dieselben Dimensionen bildet auch der im Forschungsprojekt ProwiN (Professionswissen in den Naturwissenschaften) entstandene ProwiN-Test ab (Lenske et al., 2015). Ähnlich ist auch das Verfahren von König und Blömeke (2009) strukturiert, das Aufgaben zu den Dimensionen *Strukturierung von Unterricht*, *Motivierung*, *Umgang mit Heterogenität*, *Klassenführung* und *Leistungsbeurteilung* umfasst.

Doch neben den Inhalten, die einen unmittelbaren Bezug zum Unterrichtsgeschehen aufweisen, sind auch Wissensbestände relevant, die über den Unterrichtskontext hinausgehen, wie bspw. Wissen über Beratung, Schulorganisation oder Bildungstheorie (Shulman, 1986; Voss et al., 2015). Zur Erfassung dieses konzeptuell breiter angelegten bildungswissenschaftlichen Wissens (Linninger et al., 2015) lagen zu Beginn der Testentwicklung so gut wie keine Testverfahren vor (siehe unten). Diese Lücke sollte im Rahmen des BilWiss-Projekts (Kunter et al., 2017) geschlossen werden, welches im Jahr 2009 startete. Dazu wurde im ersten Schritt eine Delphi-Befragung von Expertinnen und Experten aus der ersten und zweiten Phase der Lehrerbildung durchgeführt (Kunina-Habenicht et al., 2012). Ein wesentliches Ziel dieser Studie war die Ermittlung eines Katalogs für die relevantesten bildungswissenschaftlichen Kernthemen in der Lehrerbildung. Als zentrales Ergebnis wurden über 100 Themen aus den Bereichen Unterricht, Lernen, Entwicklung, Sozialisation, Diagnostik und Evaluation, Heterogenität und soziale Konflikte, Bildungstheorie, Bildungssystem und Schulorganisation so-

wie Lehrerberuf als zentrale bildungswissenschaftliche Wissensinhalte identifiziert.

Im nächsten Schritt wurde ein Wissenstest entwickelt, der die in der Delphi-Studie identifizierten Themen objektiv, reliabel und valide erfassen sollte. Die Besonderheit dieses Tests besteht in der Abdeckung einer Vielzahl bildungswissenschaftlicher Themen, da er die sechs Inhaltsbereiche *Unterrichtsgestaltung*, *Lernen und Entwicklung*, *Diagnostik und Evaluation*, *Bildungstheorie*, *Schule als Bildungsinstitution* und *Lehrerberuf als Profession* umfasst (Kunina-Habenicht et al., 2013; Kunter et al., 2017). Die bisherigen Befunde basieren vorwiegend auf Testdaten von Lehramtsabsolventinnen und -absolventen der Primar- und der Sekundarstufe aus dem Bundesland Nordrhein-Westfalen (NRW). Beispielaufgaben können bei Linninger et al. (2015) eingesehen werden. Der primäre Anwendungszweck lag dabei nicht auf der Individualdiagnostik, sondern darauf ein ökonomisches Forschungsinstrument zu entwickeln, welches bundesweit bei angehenden Lehrkräften in der Lehramtsausbildung und bei Lehrkräften im Beruf über alle verschiedenen Lehrämter hinweg eingesetzt werden kann – mit dem Ziel, unterschiedlichen Fragestellungen der Lehrerbildungsforschung nachzugehen. Insbesondere sollte es möglich sein, die Entwicklung des bildungswissenschaftlichen Wissens im Verlauf des Lehramtsstudiums abzubilden und dessen Erträge für den späteren Berufserfolg sichtbar zu machen. Ein weiterer langfristig anvisierter Anwendungszweck bezieht sich auf die Testanwendung im Rahmen von Evaluationsstudien und Bildungsmonitoring im Lehramtsstudium.

Mit dem BilWiss-Test, der als ökonomische Kurzversion und als inhaltlich breitere Langversion vorliegt, war es bspw. möglich, individuelle Wissensunterschiede, auch unterschiedlicher Gruppen, bezüglich des Fortschritts bzw. der Schwerpunktbildung im Lehramtsstudium abzubilden (Linninger et al., 2015; Schulze-Stocker, Holzberger, Kunina-Habenicht, Terhart & Kunter, 2016). Darüber hinaus zeigte sich, dass das mithilfe des Tests erhobene Wissen einzelne Berufserfolgsfaktoren von Lehrkräften vorhersagt, bspw. die Veränderung der Wahrnehmung des eigenen professionellen Unterrichtsverhaltens im Verlauf des Vorbereitungsdienstes (Lohse-Bossenz, Kunina-Habenicht, Dicke, Leutner & Kunter, 2015). Außerdem erweist sich bildungswissenschaftliches Wissen als Puffer gegen einen Anstieg im Beanspruchungserleben während des Vorbereitungsdienstes (Dicke et al., 2015).

Trotz dieser vielversprechenden Ergebnisse wies der BilWiss-Test noch einige psychometrische Schwächen auf. Zum einen waren insbesondere die EAP-Reliabilitäten in der Kurzform für die Skalen *Unterrichtsgestaltung* und *Schule als Bildungsinstitution* niedriger als .60. Zum anderen lieferten kognitive Interviews Hinweise auf konstrukt-irrelevante Varianz einiger Items, da für deren Lösung logisches Schließen ohne die Anwendung von Wissen als Antwortstrategie möglich war (Linninger et al., 2015). Darüber hinaus waren in der Kurzversion des BilWiss-Tests einige zentrale Delphi-Themen nicht ausreichend abgedeckt. Um diese Schwächen zu überwinden und weitreichendere Aussagen über die Validität der auf dem BilWiss-Test basierenden Testwertinterpretationen treffen zu können, wurde der bestehende BilWiss-Test einer umfangreichen Überarbeitung und Validierung unterworfen.

In der Zwischenzeit liegen neben dem BilWiss-Test weitere Verfahren vor, die (unter anderem) auch unterrichtsfernere Anteile des fachunabhängigen Wissens von Lehrkräften abdecken. Etwa zeitgleich zum Beginn des BilWiss-Projekts wurde ein weiteres Verfahren zur Erfassung des bildungswissenschaftlichen Wissens (Seifert et al., 2009; Seifert & Schaper, 2010) entwickelt, welches die Dimensionen *Erziehung und Bildung*, *Unterricht und Allgemeine Didaktik* sowie *Schulentwicklung und Gesellschaft* umfasst. Ein weiterer Test entstand vor wenigen Jahren im Rahmen des Projekts KiL (Messung professioneller Kompetenzen in mathematischen und naturwissenschaftlichen Lehramtsstudiengängen; Hohenstein, Kleickmann, Zimmermann, Köller & Möller, 2017), der neben den unterrichtsrelevanten Inhalten *Lehren*, *Lernen und Entwicklung*, *Klassenführung*, *Leistungsbeurteilung* auch die Inhaltsbereiche *Bildungssystem und Schulorganisation*, *Methoden der bildungswissenschaftlichen Forschung* sowie *Innovieren* beinhaltet. Ein Testverfahren, das sich noch in der Entwicklung befindet, ist der Essener SBW-Test (Müser, Fleischer & Leutner, 2018). Dieser Test soll die Entwicklung der Kompetenzen in den Bereichen der KMK-Standards (KMK, 2014) *Unterrichten*, *Beurteilen*, *Erziehen* und *Innovieren* im Lehramtsstudium abbilden.

Im vorliegenden Beitrag wird eine revidierte Fassung des BilWiss-Tests – der BilWiss-2.0-Test – vorgestellt, der die oben beschriebenen Probleme überwinden soll. Dieser Test ist im Forschungsprojekt BilWiss-UV („Ertrag und Entwicklung des universitären bildungswissenschaftlichen Wissens – Validierung eines Kompetenztests für Lehramtsstudierende“)[1] entstanden. Im Folgenden wird zunächst auf den Testüberarbeitungsprozess eingegangen. Im Anschluss werden Ergebnisse zur Skalierung und

[1] Dieses Forschungsvorhaben wird gefördert durch das Bundesministerium für Bildung und Forschung (BMBF) im Rahmen der Förderinitiative „Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen“ (KoKoHs) (Förderkennzeichen 01 PK15007)

zur empirischen Überprüfung der dimensionalen Struktur des neuen BilWiss-2.0-Tests dargestellt. In Anlehnung an die Arbeiten von König, Ligtvoet, Klemenz und Rothland (2017) und Tachtsoglou und König (2017) wurde der Zusammenhang des bildungswissenschaftlichen Wissens mit der Anzahl der besuchten relevanten inhaltsspezifischen Lehrveranstaltungen und Studienleistungen in Form von Noten betrachtet – als Hinweis auf die Instruktionssensitivität der Testwerte. Darüber hinaus wurden Testleistungen von Studierenden, die mindestens einmal eine Prüfung wiederholen mussten, mit Personen verglichen, die keinen Wiederholungsversuch benötigten.

Methode

Testentwicklung

Der BilWiss-2.0-Test basiert inhaltlich, wie auch die Vorgängerversion, auf der oben beschriebenen Delphi-Studie (Kunina-Habenicht et al., 2012). Zu Beginn der Optimierung fand eine kritische Analyse dieser Delphi-Themen statt, die eine Neuordnung und Zusammenfassung einiger Themen zur Folge hatte. Dies betraf vor allem den Inhaltsbereich *Unterrichtsgestaltung*, wobei darauf geachtet wurde, dass sowohl die Thematik *Merkmale von Unterrichtsqualität* (bspw. Klassenführung, kognitive Aktivierung, konstruktive Lernunterstützung) als auch die Thematik *Didaktik und Methodik im Unterricht* gleichermaßen abgebildet werden. Der optimierte Test basiert nun auf insgesamt 73 bildungswissenschaftlichen Themen, die den sechs Inhaltsbereichen *Unterrichtsgestaltung*, *Lernen und Entwicklung*, *Diagnostik und Evaluation*, *Bildungstheorie*, *Schule als Bildungsinstitution* und *Lehrerberuf als Profession* zugeordnet sind. Die vollständige Liste der abgedeckten Delphi-Themen ist zu finden in der Tabelle im ESM 1 im Anhang.

Der nächste Schritt war die psychometrische Optimierung der Testitems. Dabei lieferten sowohl Aussagen aus kognitiven Interviews zu Antwortstrategien bei der Testbearbeitung als auch Itemschwierigkeiten und Diskriminationsparameter aus Erhebungen mit der Vorgängerversion wichtige Informationen zur Qualität der einzelnen Items. Im Ergebnis wurden neue Testaufgaben konstruiert und die Aufgabenformate vereinheitlicht, sodass nun alle Testitems einfache oder komplexe Multiple-Choice-Aufgaben (Einfach- und Mehrfachauswahlaufgaben) mit jeweils vier Antwortalternativen beinhalten. Die Items zielen dabei entweder auf die Erfassung des deklarativen Wissens (Faktenwissen aus Theorien, Konzepten oder Modellen) oder konzeptuellen Wissens, welches Verknüpfungen zwischen Wissenskomponenten erfordert und auf das Verständnis über Zusammenhänge einzelner Wissenskomponenten schließen lässt.

Zur Erprobung des neuen Itempools wurden mehrere Pilotierungsuntersuchungen mit insgesamt 555 Lehramtsstudierenden und erneute kognitive Interviews durchgeführt. Anschließend wurde ein Feldtest mit 627 angehenden Lehrkräften im Vorbereitungsdienst in NRW durchgeführt. Das zentrale inhaltliche Kriterium für die finale Auswahl der Items war die Abdeckung zentraler Delphi-Themen in dem jeweiligen Inhaltsbereich. Aus psychometrischer Perspektive wurden Items mit Itemschwierigkeiten in der IRT-Metrik im Bereich zwischen -3.5 (sehr leichte Items bzw. sehr niedriges Fähigkeitsniveau der Personen) und 3.5 (sehr schwere Items bzw. sehr hohes Fähigkeitsniveau der Personen) mit ausreichend hohen Diskriminationsparametern größer gleich 0.20 (Schmidt-Atzert & Amelang, 2012) berücksichtigt. Ein weiteres zusätzliches Kriterium bildeten die Ergebnisse der Analyse der Antwortstrategien aus den kognitiven Interviews, indem auffällige Items, bei denen die richtige Antwort durch logisches Schließen hergeleitet werden konnte, entweder eliminiert oder nochmal überarbeitet wurden.

Der aktuelle Test erlaubt mit insgesamt 119 Items eine umfassende Erfassung des bildungswissenschaftlichen Wissens. Darüber hinaus wurde eine Kurzversion des Tests mit 65 Items entwickelt, die es ermöglicht, zentrale Inhalte aus allen sechs Inhaltsbereichen innerhalb einer Erhebung mit einer durchschnittlichen Erhebungszeit von unter einer Zeitstunde zu erfassen. Bei der Erstellung der Kurzskalen für die einzelnen Inhaltsbereiche wurde inhaltlich Augenmerk auf die Abdeckung zentraler Delphi-Themen gelegt. Aus psychometrischer Sicht wurden Items mit einer ausreichend hohen Diskrimination (größer gleich 0.20) gewählt. Um sicherzustellen, dass der BilWiss-2.0-Kurztest in allen Fähigkeitsbereichen möglichst gut diskriminiert, wurden die Schwierigkeitsparameter der Items so gewählt, dass nach Möglichkeit der gesamte Leistungsbereich in der IRT-Metrik zwischen -3.5 und 3.5 durch die jeweilige Kurzskala abgedeckt wurde. Der hier vorliegende Beitrag berichtet Ergebnisse zur kurzen und langen Version des BilWiss-2.0-Tests. Für den Test liegt ein Testmanual mit Hinweisen zur Testnutzung vor. Die aktuelle Testfassung ist zusammen mit den Nutzungshinweisen verfügbar auf Nachfrage bei den Autorinnen und Autoren.

Studiendesign und Stichprobe

Die Datenerhebung fand im Sommersemester 2017 an vier Universitäten in Essen, Frankfurt, München und Tübingen statt. Aus der Erhebung liegen Daten von insgesamt 928 Personen vor. Nach dem Ausschluss der Teil-

nahmen mit unzureichender Bearbeitungsqualität[2] bildet eine Stichprobe von $N = 788$ Personen aus den vier Erhebungsstandorten (Essen: $n = 234$, Frankfurt: $n = 384$, München: $n = 122$ und Tübingen: $n = 48$) die Grundlage für alle Analysen. Davon haben 25 Personen aus Essen den Wissenstest auf Papier beantwortet, die übrigen 763 Personen haben die Umfrage online bearbeitet. 553 Personen (70.2% der Studierenden) waren weiblich. Das durchschnittliche Alter betrug 22.5 Jahre ($SD = 3.5$; Range: 18–46). Die Studierenden verteilen sich auf das Lehramt an Grundschulen (14.1%), an Haupt- und Realschulen (15.1%), an Gymnasien und Gesamtschulen (46.8%), an Förderschulen (7.7%) und an beruflichen Schulen (16.0%).

Instrumente

Aufgrund der großen Itemanzahl wurde ein Rotationsdesign mit vier Testheften verwendet, wobei der BilWiss-2.0-Kurztest in jedem Testheft vollständig enthalten war. Neben dem bildungswissenschaftlichen Wissen wurden unter anderem demographische Angaben zu Geschlecht, Alter und Lehramtszugang erfasst.

Zur Ermittlung des standortübergreifenden Studienfortschritts im bildungswissenschaftlichen Teilstudium haben die Studierenden angegeben, welche der vorgegebenen Module sie im Rahmen ihres bildungswissenschaftlichen Studiums bereits absolvieren konnten. Der Studienfortschritt wurde über die Anzahl der bisher belegten Semesterwochenstunden (SWS) und erhaltenen Credit Points (CP) ermittelt. Zusätzlich wurden die Modulhandbücher für die vier Standorte daraufhin analysiert, inwieweit die sechs Inhaltsbereiche des Tests durch die Module abgedeckt werden. Die Kodierung erfolgte auf der Grundlage der inhaltlichen Passung der Testaufgaben und Modulbeschreibungen durch zwei Kodierer, die sich auf folgende Kodierung pro Modul und Studiengang einigten: 1 – leichter Wissenszuwachs im Inhaltsbereich erwartbar (5–7 Items der Teilskala werden durch das Modul abgedeckt) bzw. 2 – Wissenszuwachs im Inhaltsbereich wahrscheinlich (8 oder mehr Items). Auf der Grundlage dieser Kodierung wurde dann der Umfang der relevanten besuchten Lehrveranstaltungen für die sechs Inhaltsbereiche als CP und SWS berechnet. Es wurde sichergestellt, dass die Kodiererinnen und Kodierer sich in allen Fällen auf eine eindeutige Kodierung pro Modul und Studiengang einigen konnten.

Als ein Indikator für Studienerfolg wurde erfragt, ob eine Modulprüfung in der Vergangenheit einmal oder mehrmals wiederholt werden musste. Zudem wurden ein Semester später bisherige Studiennoten in den Bildungswissenschaften erfragt, die innerhalb der Standorte z-standardisiert wurden, um die Vergleichbarkeit der Noten zwischen den Standorten zu gewährleisten. Dabei stehen kleinere Notenwerte für bessere Studienleistungen.

Statistische Auswertung

Die IRT-basierte Schätzung der Itemschwierigkeiten und Diskriminationsparameter erfolgte für den gesamten Kurztest und separat für jede der sechs Testskalen mit eindimensionalen 2-PL-Partial-Credit-Modellen (Muraki, 1992) im R-Paket *TAM* (Robitzsch, Kiefer & Wu, 2017), jeweils getrennt für die kurze und lange Version. Auf dieser Grundlage wurden die EAP/PV-Reliabilität[3] (für weitere Erläuterungen siehe Lohse-Bossenz et al., 2015) sowie EAP-Personenparameter geschätzt. Zusätzlich wurde die Testinformation für die Kurz- und Langfassung der Teilskalen bestimmt. Detaillierte Ergebnisse der IRT-Analysen umfassen insbesondere die deskriptive Verteilung der Itemschwierigkeiten und Personenparameter (ESM 3 und ESM 4) und Testinformationskurven für die einzelnen Inhaltsbereiche im Kurz- und Langtest (ESM 5).

Im IRT-Kontext stellen die Beurteilung der Modellpassung für 2-PL-Partial-Credit-Modelle und die Schätzung von mehrdimensionalen Modellen mit mehr als drei Dimensionen mit traditionellen Maximum-Likelihood-Schätzverfahren eine methodische Herausforderung dar. Zur Überprüfung der empirischen Struktur des Konstrukts *Bildungswissenschaftliches Wissen* wurden daher Strukturgleichungsmodelle (SEM) für den Kurztest in der Software Mplus 8 (Muthén & Muthén, 1998–2017) mit dem Weighted Least Squares Mean Variance- (WLSMV) Schätzer geschätzt. Diese Prozedur wird empfohlen für Schätzungen von SEM mit kategorialen bzw. ordinalen Daten (Kline, 2011). Zur Beurteilung der Modellpassung wurden die traditionellen Cut-Off-Werte für die Modellfit-Indizes Root Mean Square Error of Approximation (RMSEA) und Confirmatory Fit Index (CFI) herangezogen. Eine gute Modellpassung ist angezeigt durch RMSEA-Werte kleiner als .05 und CFI-Werte größer als .95, während RMSEA-Werte kleiner als .08 und CFI-Werte größer als .90 eine akzeptable Modellpassung indizieren (Hu & Bentler, 1999;

[2] Die Bearbeitungsqualität wurde als unzureichend klassifiziert, wenn die Bearbeitungszeit der Testaufgaben bei mehr als der Hälfte der bearbeiteten Umfrageseiten so gering war, dass eine ernsthafte Bearbeitung der Aufgaben praktisch unmöglich war.

[3] Die EAP/PV (Expected-A-Posteriori/Plausible Values) Reliabilität gibt den Anteil der modellbasiert geschätzten aufgeklärten Varianz im Verhältnis zur Gesamtvarianz an. Eine exakte Beschreibung der verwendeten R-Funktion „WLErel" im TAM-Paket ist zu finden unter https://cran.r-project.org/web/packages/TAM/TAM.pdf

Tabelle 1. Deskriptive Statistiken für die Personenparameter in der Kurz- und Langform des BilWiss-2.0-Tests

	Kurztest			Langtest		
Inhaltsbereich	*SD*	Min	Max	*SD*	Min	Max
Unterrichtsgestaltung	0.84	-3.30	2.25	0.85	-3.73	2.34
Lernen und Entwicklung	0.70	-2.50	2.17	0.72	-2.96	2.17
Diagnostik und Evaluation	0.76	-2.60	2.67	0.77	-2.61	2.65
Bildungstheorie	0.80	-2.59	1.93	0.80	-2.60	1.93
Schule als Bildungsinstitution	0.77	-2.48	2.47	0.79	-2.85	2.47
Lehrerberuf als Profession	0.77	-2.80	1.91	0.78	-3.19	2.02
Kurztest gesamt	0.94	-4.69	2.37			

Anmerkungen: Die Kennwerte beziehen sich auf die geschätzten IRT-Personenparameter in der Kurz- und Langform des BilWiss-2.0-Tests. Um die Identifizierbarkeit des Modells sicherzustellen, wurde der Populationsmittelwert auf 0 fixiert. *SD:* Standabweichung; Min: Minimum; Max: Maximum

für kategoriale Daten Xia & Yang, 2018; Yu, 2002). Eine Schätzung des analogen Modells für den Langtest ist nicht möglich, da für die Langform des BilWiss-2.0-Tests keine vollständige Kovarianzmatrix vorliegt. Die Schätzung des sechsdimensionalen IRT-Modells war aufgrund der Komplexität des zugrunde liegenden Schätzverfahrens nicht realisierbar. Omega (latente interne Konsistenz) wurde nach McDonald (1999) auf der Grundlage der standardisierten Faktorladungen im sechsfaktoriellen Messmodell bestimmt.

Ergebnisse

IRT-Skalierung

In Tabelle 1 und ESM 2 sind die Ergebnisse der IRT-Skalierung zusammengefasst. Die geschätzten Personenparameter des Kurz- und Langtests für die einzelnen Inhaltsbereiche korrelieren zwischen r = .95 und r = 1. In Tabelle 1 sind deskriptive Statistiken für die Personenparameter im Kurz- und Langtest dargestellt. Im ESM 2 sind deskriptive Kennwerte zur Verteilung der Itemschwierigkeiten und Diskriminationsparameter aufgeschlüsselt nach Inhaltsbereichen angegeben (siehe auch Histogramme in ESM 3 und ESM 4).

Im Langtest ist die EAP-Reliabilität für den Inhaltsbereich *Unterrichtsgestaltung* zufriedenstellend (.72), und für die meisten anderen Inhaltsbereiche größer oder gleich .60 und damit noch akzeptabel (Tabelle 2). Eine Ausnahme bildet der Inhaltsbereich *Lernen und Entwicklung*, der eine zu geringe EAP-Reliabilität aufweist. Im Kurztest sind die EAP-Reliabilitäten generell etwas niedriger als im Langtest und liegen nur für die Hälfte der Skalen im zufriedenstellenden bis akzeptablen Bereich (größer oder gleich .60).

Vergleicht man die Testinformationskurven zwischen Kurz- und Langtest (siehe ESM 5 im Anhang), so wird deutlich, dass die Testinformation für alle Teilskalen im Langtest erwartungskonform höher ist als in der Kurzfassung. Die höchste Testinformation liegt für den Bereich Unterrichtsgestaltung vor, was sich auch in der zufriedenstellenden EAP-Reliabilität widerspiegelt, da die Testinformation in die Berechnung der EAP-Reliabilität eingeht. Die Testinformation im Kurztest für die Teilskalen *Lernen und Entwicklung* und *Diagnostik und Evaluation* ist hingegen im Vergleich zu den übrigen Skalen auffällig niedrig, was sich auch in der damit einhergehenden unzureichenden EAP-Reliabilität zeigt.

Im nächsten Schritt wurde die Reliabilität für den gesamten Kurztest betrachtet. Dabei zeigte sich, dass sowohl die interne Konsistenz für den Kurztest (Cronbachs α = 0.86) als auch die EAP-Reliabilität im entsprechenden 2-PL-Partial-Credit-Modell mit 0.88 (siehe auch Tabelle 2) sehr zufriedenstellend waren.

Hinweise auf die Validität der Testwerte

Überprüfung der faktoriellen Struktur des Kurztests

Zur Überprüfung der faktoriellen Struktur des Kurztests und zur Bestimmung der latenten Reliabilität (Omega) wurde ein SEM mit sechs latenten korrelierten Faktoren geschätzt. Der Modellfit ist gut (χ^2 = 2271.1; df = 2000; p < .001, CFI = .958; RMSEA = 0.013). Dieses Modell weist einen deutlich besseren Modellfit auf als das eindimensionale SEM mit einem Generalfaktor über alle Kurztestitems (χ^2 = 2420.6; df = 2015; p < .001, CFI = .938; RMSEA = 0.016). Die latenten Korrelationen zwischen den sechs Inhaltsbereichen liegen zwischen .55 und .93 und sind alle statistisch signifikant (vgl. Tabelle 3). Besonders hohe Korrelationen weisen die Inhaltsbereiche *Lehrerberuf als Profession* und *Schule als Bildungsinstitution* (r = .93) bzw. *Unterrichtsgestaltung* und *Lernen und Ent-*

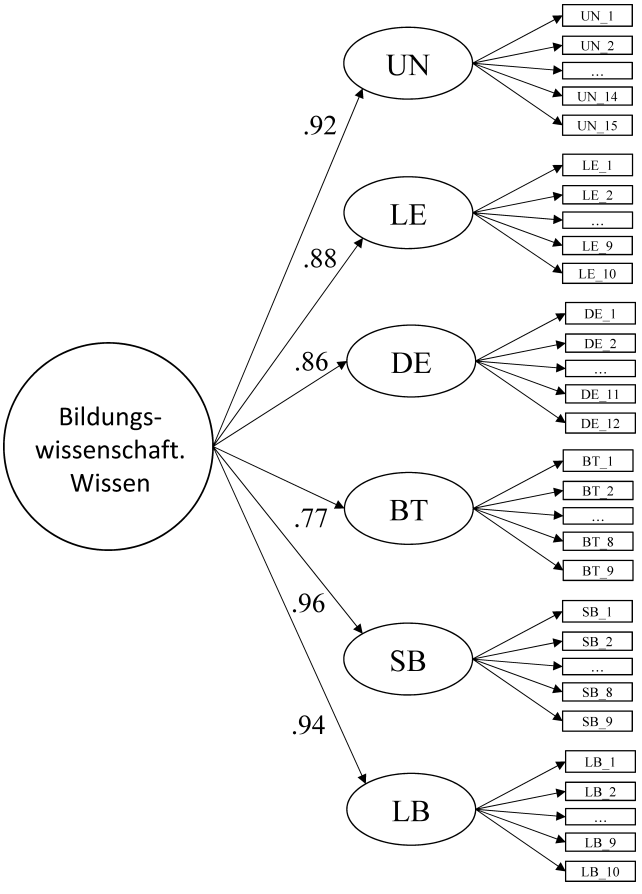
Tabelle 2. Reliabilität der Skalen für die Kurz- und Langform des BilWiss-2.0-Tests

Inhaltsbereich	Anzahl Items Kurztest	Anzahl Items Langtest	EAP-Reliabilität Kurztest	EAP-Reliabilität Langtest	Omega[1] Kurztest
Unterrichtsgestaltung	15	23	.71	.72	.76
Lernen & Entwicklung	10	24	.49	.52	.53
Diagnostik & Evaluation	12	21	.57	.59	.55
Bildungstheorie	9	14	.64	.65	.68
Schule als Bildungsinstitution	9	19	.60	.62	.65
Lehrerberuf als Profession	10	18	.59	.61	.66
Kurztest gesamt	65		.88		
Gesamt	65	119			

Anmerkungen: [1] Omega (latente interne Konsistenz) wurde berechnet nach McDonald (1999) auf der Grundlage des Strukturgleichungsmodells mit sechs korrelierten Faktoren.

wicklung (*r* = .92) auf. Eine mögliche Erklärung für die hohen Zusammenhänge zwischen diesen drei Faktoren besteht darin, dass alle drei Inhaltsbereiche unterrichtsnahe bzw. schulpraktische Themen umfassen.

Da die Korrelationen zwischen den Inhaltsbereichen zwar hoch, jedoch kleiner als eins sind, wurde im zweiten Schritt ein – sparsameres und inhaltlich einfacher interpretierbares – hierarchisches SEM mit sechs latenten Faktoren erster Ordnung und einem Generalfaktor zweiter Ordnung für bildungswissenschaftliches Wissen geschätzt (Abbildung 1). Das Modell weist einen guten Modellfit auf (χ^2 = 2311.3; *df* = 2009; $p < .001$; CFI = 0.954; RMSEA = 0.014). Es muss allerdings angemerkt werden, dass in beiden SEM die Ladungen auf die Faktoren erster Ordnung für einen Leistungstest vergleichsweise niedrig sind: Je nach Inhaltsbereich variieren sie zwischen .07 und .61; zwei der 65 Items weisen statistisch nichtsignifikante Ladungen auf, was sich auch in der zu niedrigen latenten internen Konsistenz für die Skalen *Diagnostik und Evaluation* und *Lernen und Entwicklung* niederschlägt. Omega – als Indikator für latente interne Konsistenz – ist für alle Teilskalen bis auf *Diagnostik und Evaluation* etwas höher als die EAP-Reliabilität (vgl. Tabelle 2).



Anmerkungen: UN: Unterrichtsgestaltung; LE: Lernen & Entwicklung; DE: Diagnostik & Evaluation; BT: Bildungstheorie; SB: Schule als Bildungsinstitution; LB: Lehrerberuf als Profession. Modell Fit: χ^2 = 2311.3; df = 2009; $p < .001$; CFI = .954; RMSEA = 0.014.

Abbildung 1. Schematische Darstellung des hierarchischen Strukturgleichungsmodells für den BilWiss-2.0-Kurztest.

Korrelationen mit der Anzahl der relevanten besuchten inhaltsspezifischen Lehrveranstaltungen und Studienleistungen und Mittelwertvergleiche für bekannte Gruppen

In Tabelle 4 sind die Korrelationen zwischen den Personenparametern der einzelnen Inhaltsbereiche und der Anzahl der relevanten besuchten inhaltsspezifischen Lehrveranstaltungen (in CP und SWS) und Studienleistungen dargestellt. Erwartungsgemäß korrelieren die Anzahl der CP und SWS sehr hoch miteinander (*r* zwischen .95 und .98). Bezüglich des Zusammenhangs mit der Anzahl der relevanten besuchten inhaltsspezifischen Lehrveranstaltungen zeigt insbesondere die Anzahl der bereichsspezifischen CP (mit der Ausnahme der Skala *Bildungstheorie)* statistisch signifikante positive Korrelationen mit den

Tabelle 3. Latente Korrelationen zwischen den sechs Dimensionen im Strukturgleichungsmodell mit sechs korrelierten latenten Faktoren für die Kurzform des BilWiss-2.0-Tests

Inhaltsbereich	UN	LE	DE	BT	SB
Unterrichtsgestaltung (UN)	1				
Lernen und Entwicklung (LE)	.92	1			
Diagnostik und Evaluation (DE)	.78	.71	1		
Bildungstheorie (BT)	.70	.55	.71	1	
Schule als Bildungsinstitution (SB)	.81	.78	.88	.80	1
Lehrerberuf als Profession (LB)	.87	.86	.74	.70	.93

Anmerkungen: Alle Korrelationen sind statistisch signifikant ($p < .05$). Der Modell-Fit des Modells war gut ($\chi^2 = 2271.1$; df = 2000; $p < .001$; CFI = .958; RMSEA = 0.013).

Tabelle 4. Korrelationen zwischen der Leistung im BilWiss-2.0-Test und der Anzahl der relevanten besuchten inhaltsspezifischen Lehrveranstaltungen und Studienleistungen im bildungswissenschaftlichen Studium

Inhaltsbereich	Inhaltsspezifische CP	Inhaltsspezifische SWS	Durchschnittliche Note
Kurztest			
Unterrichtsgestaltung	.23*	.22*	-.20*
Lernen und Entwicklung	.10*	.07	-.12*
Diagnostik und Evaluation	.16*	.16*	-.19*
Bildungstheorie	.03	.01	-.23*
Schule als Bildungsinstitution	.16*	.11*	-.27*
Lehrerberuf als Profession	.11*	.11*	-.22*
Langtest			
Unterrichtsgestaltung	.24*	.22*	-.21*
Lernen und Entwicklung	.12*	.09	-.14*
Diagnostik und Evaluation	.17*	.17*	-.22*
Bildungstheorie	.03	.02	-.23*
Schule als Bildungsinstitution	.14*	.10*	-.26*
Lehrerberuf als Profession	.11*	.10*	-.21*
Kurztest gesamt[1]	.23*	.21*	-.29*

Anmerkungen: CP: Credit points; SWS: Semesterwochenstunden; statistisch signifikante Korrelationen ($p < .05$) sind markiert mit *. [1] Korrelation zwischen SEM-Generalfaktor für bildungswissenschaftliches Wissen und Gesamtanzahl belegter CP und SWS in Bildungswissenschaften

Personenparametern der einzelnen Testskalen. Zusätzlich wurden in weiteren SEMs Korrelationen zwischen dem Generalfaktor für bildungswissenschaftliches Wissen und den bisher belegten CP und SWS in den Bildungswissenschaften geschätzt. Diese sind ebenso erwartungskonform statistisch signifikant ($r = .23$ für CPs und $r = .21$ für SWS). Darüber hinaus sind alle Korrelationen mit Studiennoten statistisch signifikant (r zwischen -.12 und -.27). Auch wenn die gefundenen Zusammenhänge eher klein sind, liefern sie dennoch wichtige Hinweise für konvergente und prognostische Validität der Interpretationen der Testwerte im revidierten Test.

Weitere Belege für die Validität der Testwertinterpretationen liefern Mittelwertvergleiche zwischen Studierenden, die mindestens einmal eine Modulprüfung wiederholen mussten, und Studierenden, die alle bisherigen Prüfungen beim ersten Versuch bestanden haben. Dabei zeigte sich für alle Inhaltsbereiche, dass die zweite Gruppe statistisch signifikant bessere Testleistungen mit moderaten bis mittelgroßen Effektstärken aufweist (siehe Tabelle 5). Die größten Effektstärken wurden für die Inhaltsbereiche Unterrichtsgestaltung ($d = .51$ für Kurztest und $d = .54$ für Langtest) und Schule als Bildungsorganisation ($d = .54$ für Kurztest und $d = .56$ für Langtest) identifiziert.

Diskussion

Ziel dieses Beitrags war es, einen revidierten Test zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften, den BilWiss-2.0-Test, vorzustellen

Tabelle 5. Vergleich der Wissensleistungen im Kurz- und Langtest für Personen, die zum befragten Messzeitpunkt mindestens einmal (N = 103) vs. kein einziges Mal durch eine Modulprüfung durchgefallen sind (N = 675)

Inhaltsbereich	*M* (*SD*) durchgefallen (*N* = 103)	*M* (*SD*) nicht durchgefallen (*N* = 675)	*t(df)*	*p*	*d*
Kurztest					
Unterrichtsgestaltung	-.36 (.89)	.06 (.82)	-4.82 (776)	< .001	.51
Lernen und Entwicklung	-.19 (.70)	.03 (.69)	-3.08 (776)	.002	.33
Diagnostik und Evaluation	-.18 (.74)	.03 (.76)	-2.62 (776)	.009	.28
Bildungstheorie	-.26 (.80)	.04 (.79)	-3.53 (776)	< .001	.37
Schule als Bildungsinstitution	-.36 (.64)	.05 (.77)	-5.93 (152)	< .001	.54
Lehrerberuf als Profession	-.25 (.76)	.04 (.76)	-3.56 (776)	< .001	.38
Langtest					
Unterrichtsgestaltung	-.39 (.91)	.06 (.82)	-5.12 (776)	< .001	.54
Lernen und Entwicklung	-.24 (.74)	.04 (.70)	-3.68 (776)	< .001	.39
Diagnostik und Evaluation	-.20 (.71)	.03 (.77)	-2.80 (776)	.005	.30
Bildungstheorie	-.26 (.80)	.04 (.79)	-3.57 (776)	< .001	.38
Schule als Bildungsinstitution	-.37 (.68)	.06 (.78)	-5.88 (147)	< .001	.56
Lehrerberuf als Profession	-.27 (.79)	.04 (.77)	-3.85 (776)	< .001	.41
Kurztest gesamt*	58.75 (13.24)	66.17 (13.61)	-5.18 (776)	< .001	.55

Anmerkungen: Mittelwertvergleiche wurden berechnet mittels *t*-Test bzw. Welch-Test, falls die Annahme der Varianzhomogenität verletzt war. Alle Vergleiche waren statistisch signifikant ($p < .05$). *M:* Mittelwert; SD: Standardabweichung; *t: t*-Wert; *df:* Anzahl der Freiheitsgrade; *p:* exakter p-Wert; *d:* Effektstärke.
* Es handelt sich um den manifesten Gesamtscore im Kurztest.

und Hinweise auf die psychometrische Güte der mit dem Instrument gemessenen Testwerte zu präsentieren. Zentrales Ziel war hierbei die Entwicklung eines Forschungsinstruments, welches unter anderem die Entwicklung des bildungswissenschaftlichen Wissens im Verlauf des Lehramtsstudiums adäquat abbildet.

Durch die Testüberarbeitung ist es im Vergleich zur Vorgängerversion gelungen – insbesondere auch im Kurztest – eine Abdeckung der in der vorangegangenen Delphi-Studie (Kunina-Habenicht et al., 2012) als am relevantesten eingeschätzten Themen zu erreichen. Die Ergebnisse der Strukturgleichungsmodellierung zeigen, dass sich mithilfe des Tests Subdimensionen des bildungswissenschaftlichen Wissens abbilden lassen, die allerdings zum Teil hoch miteinander korreliert sind. Entsprechend kann es als sinnvoll erachtet werden, die Subdimensionen zu einer Dimension höherer Ordnung zusammenzufassen. Ferner wurden in der hier vorliegenden Erhebung – im Gegensatz zu bisherigen Studien, bei denen der BilWiss-Test zum Einsatz kam – auch das Lehramt an Förderschulen und das Lehramt an beruflichen Schulen einbezogen. Eine weitere Stärke der hier vorliegenden Studie ist die differenzierte und genauere Erfassung der inhaltsspezifischen Lerngelegenheiten im bildungswissenschaftlichen Studium.

Für die Validität der Testwertinterpretationen der revidierten Testfassung sprechen die erwartungskonformen statistisch signifikanten – wenn auch niedrigen – Korrelationen zwischen der Testleistung und der Anzahl der relevanten besuchten inhaltsspezifischen Lehrveranstaltungen und Studienleistungen sowie statistisch signifikante Unterschiede in den Wissensleistungen zwischen Studierenden, die mindestens einmal eine Modulprüfung wiederholen mussten, und Personen, die keinen Wiederholungsversuch benötigt hatten. Die Größe der gefundenen Effekte entspricht den Befunden, die typischerweise in der Literatur berichtet werden (z. B. Tachtsoglou & König, 2017). Es ist fraglich, ob man höhere Korrelationen zwischen Testleistungen und Studiennoten erwarten kann, denn letztlich spiegeln die analysierten Modulhandbücher nur das intendierte Curriculum wider (Schulze-Stocker, Holzberger & Lohse-Bossenz, 2017). Folglich kann man nicht empirisch beurteilen, wie die Modulbeschreibung in den Vorlesungen und Seminaren tatsächlich von den Lehrenden an den Hochschulen umgesetzt wurde (implementiertes Curriculum) und welches Wissen Studierende tatsächlich aktiv aufgebaut haben. Daher interpretieren wir die dargestellten Befunde dahingehend, dass der BilWiss-2.0-Test instruktionssensitiv ist und interindividuelle Unterschiede im Verlauf des Lehramtsstudiums in erwarteter Weise abbildet.

In Bezug auf die Reliabilität der Testskalen liegen die Werte für die Langskalen vorwiegend und für die Kurzskalen zur Hälfte im akzeptablen Bereich und sind vergleichbar mit den Reliabilitäten anderer Tests zur Erfassung des pädagogischen bzw. bildungswissenschaftlichen

Wissens (Hohenstein et al., 2017; König & Blömeke, 2009). König und Blömeke (2009) berichten bspw. für eine ähnlich gemischte Studierendenstichprobe unterschiedlicher Universitäten in Deutschland ebenfalls unzureichende interne Konsistenzen für einige Testskalen zwischen .52 und .59. Man kann argumentieren, dass es bei der Konstruktion von Wissenstests generell relativ schwer ist, ähnlich hohe Reliabilitäten wie bspw. bei fluiden Intelligenztests zu erreichen, da Intelligenztests in der Regel Untertests mit schematisch sehr ähnlichen Aufgaben beinhalten. Dagegen ist es bei der Abfrage des Wissens aus umfangreichen Themengebieten – wie es auch beim BilWiss-Wissenstest der Fall ist – fraglich, ob austauschbare Aufgaben überhaupt konstruiert werden können, so dass bei Wissenstests vermutlich nahezu immer eine größere thematische Heterogenität auftritt, die sich in niedrigeren internen Konsistenzen niederschlägt. Die thematische Breite des bildungswissenschaftlichen Wissens führt darüber hinaus dazu, dass die latente Variable in den vorliegenden Modellierungen nicht als die für die Lösung verursachende Variable verstanden, sondern eher als Hilfsmittel zur Aggregation von Informationen aus dem jeweiligen Inhaltsbereich verwendet wird.

Ein weiterer Faktor, der sich negativ auf die Reliabilität der Testskalen auswirken könnte, betrifft die Studierendenstichprobe mit sehr heterogenem Studienfortschritt. In der Studie von König und Blömeke (2009) verbesserten sich die Reliabilitäten für dieselben Skalen, wenn ausschließlich die im Studium fortgeschrittenen Studierenden berücksichtigt werden (auf .63–.64). Dies spricht dafür, dass zu einem frühen Zeitpunkt im Studium lediglich „verinseltes Wissen“ statt einer kohärenten Wissensbasis aufgebaut werden konnte (König & Blömeke, 2009; vgl. hierzu auch Oser, 2001).

Obwohl die angestrebten Verbesserungen der psychometrischen Kennwerte der Kurzskalen der Inhaltsbereiche *Unterrichtsgestaltung* und *Schule als Bildungsinstitution* gelungen sind, müssen wir gleichzeitig jedoch eine ungünstige Entwicklung der Reliabilitäten anderer Testskalen feststellen, vor allem bei der Skala *Lernen und Entwicklung*. Dies kann zum Teil durch die deutliche Verringerung der Itemanzahl erklärt werden. So waren bspw. für den Inhaltsbereich *Lernen und Entwicklung* 10 Items in der Kurzform und 24 Items in der Langform im BilWiss-2.0-Test vs. 15 Items im Kurztest und 69 Items im Langtest im ursprünglichen BilWiss-Test enthalten (Linninger et al., 2015). Diese starke Reduktion der Itemanzahl bei möglichst guter Delphi-Themenabdeckung führte bei der Mehrzahl der Testskalen dazu, dass in der Langform weniger Items je Delphi-Thema berücksichtigt werden konnten, so dass ein ohnehin heterogenes Konstrukt nun mit weniger Items gemessen wird, was sich negativ auf die Reliabilität (insbesondere latente interne Konsistenz) auswirken kann. Auch in Bezug auf die neue Zusammenstellung der Items in der Kurzform des BilWiss-2.0-Tests, bei der eine höhere Delphi-Abdeckung realisiert wurde, ist zu befürchten, dass die Heterogenität der abgedeckten Inhalte in den Inhaltsbereichen verstärkt wurde, was ein möglicher Grund für die Verschlechterung der EAP-Reliabilitäten in vier von sechs Kurzskalen ist. Wie in Tabelle ESM 1 zu sehen, ist bspw. jedes Delphi-Thema innerhalb des Inhaltsbereichs *Lernen und Entwicklung* in der Langfassung bis auf zwei Ausnahmen jeweils nur durch ein Item vertreten. Zusätzlich zur geringen Anzahl der Items pro Thema ist die Heterogenität der Delphi-Themen insbesondere innerhalb des Inhaltsbereichs *Lernen und Entwicklung* ein weiterer wahrscheinlicher Grund für die unbefriedigende EAP-Reliabilität der Skala, da dieser Inhaltsbereich ein breites Spektrum von entwicklungspsychologischen, soziologischen und lerntheoretischen Themen umfasst.

Die hier vorgestellten Ergebnisse zum BilWiss-2.0-Test machen deutlich, dass mit diesem revidierten Verfahren sechs unterschiedliche – wenn auch zum Teil hoch untereinander korrelierende – Inhaltsbereiche des bildungswissenschaftliches Wissen erfasst werden können, wobei sich die Erhöhung der Testökonomie in dieser revidierten Testfassung bei einigen Inhaltsbereichen zulasten der Reliabilität der Skalen auswirkt. Positiv hervorzuheben ist, dass es im Zuge der Revision insbesondere gelungen ist die Reliabilität der Lang- und Kurzskala für den Inhaltsbereich *Unterrichtsgestaltung* zu verbessern. Darüber hinaus ist es sehr erfreulich, dass sich der Kurztest als Gesamttest als sehr reliabel erwies und eine hohe interne Konsistenz bzw. hohe EAP-Reliabilität aufwies.

Ein zentrales Anliegen im BilWiss-2.0-Test war es, eine adäquate Breite an bildungswissenschaftlichen Themen in der Lehramtsbildung abzubilden. Die Inhalte der Testitems orientieren sich daher an den relevanten bildungswissenschaftlichen Themen, die in der vorangegangenen Delphi-Studie ermittelt wurden. Wir befürchten, dass man der Breite des theoretischen Konstrukts nicht gerecht werden würde, wenn man den Kurztest noch weiter kürzen würde. Aus unserer Sicht besteht bei der Testkonstruktion häufig die Kunst darin, die schwierige Balance herzustellen zwischen dem messtheoretisch angestrebtem Ziel, eine möglichst inhaltlich homogene Skala zu konstruieren, und dem Anspruch, ein inhaltlich breites Konstrukt möglichst sparsam zu erfassen. Dabei entsteht in unserem Fall ein Trade-off zwischen der inhaltlichen Abdeckung der Delphi-Themen und psychometrischen Gütekriterien.

Im Gegensatz zu einigen Langskalen des Wissenstests, die teilweise niedrige Diskriminationsparameter aufweisen, sprechen die psychometrischen Kennwerte für die Verwendung des Kurztests als globaler Indikator für bildungswissenschaftliches Wissen sowie für die Verwendung der Kurz- und Langskala für den Inhaltsbereich *Un-*

terrichtsgestaltung. Aufgrund der zum Teil niedrigen EAP-Reliabilitäten für die Langskalen in einigen anderen Inhaltbereichen sollte dieses Instrument nicht für Individualdiagnostik eingesetzt werden. Weitere anstehende Datenerhebungen im Validierungsprogramm des Forschungsprojekts BilWiss-UV, wie bspw. die längsschnittliche Verfolgung der hier vorgestellten Studierendenstichprobe, werden zusätzliche Hinweise auf die psychometrische Güte der Testwerte und deren Optimierungspotenzial geben. Darüber hinaus liegen erste Hinweise zur curricularen Validität der Testwerte im BilWiss-2.0-Test aus dem Abgleich der Delphi-Themen mit KMK-Standards und Expertenratings der Ländervertreterinnen und -vertretern außerhalb von NRW vor. Diese Ergebnisse sprechen für die länderübergreifende Nutzung des BilWiss-2.0-Tests als Forschungsinstrument (Kunina-Habenicht, Maurer, Schulze-Stocker, Wolf, Hein, Leutner, Seidel & Kunter, 2019).

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version dieses Artikels verfügbar unter https://doi.org/10.1026/0012-1924/a000238

ESM 1. Übersicht über die berücksichtigten Delphi-Themen in der Kurz- und Langform des BilWiss-2.0-Tests

ESM 2. Verteilung der Itemkennwerte in der IRT-Skalierung für die Kurz- und Langform des BilWiss-2.0-Tests

ESM 3. Deskriptive Verteilung der IRT-Personenparameter und Itemschwierigkeiten für den Kurztest und die verschiedenen Wissensdimensionen für die Kurzform des BilWiss-2.0-Tests (N = 788)

ESM 4. Deskriptive Verteilung der IRT-Personenparameter und Itemschwierigkeiten in verschiedenen Wissensdimensionen für die Langform des BilWiss-2.0-Tests (N = 788)

ESM 5. Testinformationskurven für den Kurztest sowie für die Kurz- und Langform der einzelnen Inhaltsbereiche im BilWiss-2.0-Tests

Literatur

Alles, M., Apel, J., Seidel, T. & Stürmer, K. (2019). Candidate Teachers Experience Coherence in University Education and Teacher Induction: the Influence of Perceived Professional Preparation at University and Support during Teacher Induction. *Vocations and Learning, 12*, 87 – 112. https://doi.org/10.1007/s12186-018-9211-5

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. et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47,* 133 – 180.

Blömeke, S., Bremerich-Vos, A., Kaiser, G., Nold, G. & Schwippert, K. (2013). Kompetenzen im Studienverlauf: Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrerausbildung aus TEDS-LT. Münster: Waxmann.

Darling-Hammond, L., Chung, R. & Frelow, F. (2002). Variation in Teacher Preparation: How Well Do Different Pathways Prepare Teachers To Teach. *Journal of Teacher Education, 53*, 286 – 302.

Dicke, T., Parker, P. D., Holzberger, D., Kunina-Habenicht, O., Kunter, M. & Leutner, D. (2015). Beginning teachers' efficacy and emotional exhaustion: Latent changes, reciprocity, and the influence of professional knowledge. *Contemporary Educational Psychology, 41*, 62 – 72.

Hohenstein, F., Kleickmann, T., Zimmermann, F., Köller, O. & Möller, J. (2017). Erfassung von pädagogischem und psychologischem Wissen in der Lehramtsausbildung: Entwicklung eines Messinstruments. *Zeitschrift für Pädagogik, 63* (1), 91 – 113.

Hohenstein, F., Zimmermann, F., Kleickmann, T., Köller, O. & Möller, J. (2014). Sind die bildungswissenschaftlichen Standards für die Lehramtsausbildung in den Curricula der Hochschulen angekommen? *Zeitschrift für Erziehungswissenschaft, 17*, 497 – 507.

Hu, L. T. & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling, 6*, 1 – 55.

Jüttner, M., Boone, W., Park, S. & Neuhaus, B. J. (2013). Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). *Educational Assessment, Evaluation and Accountability, 25* (1), 45 – 67.

Kleickmann, T., Großschedl, J., Harms, Z., Heinze, A., Herzog, S., Hohenstein, F. et al. (2014). Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer- Testentwicklung im Rahmen des Projekts KiL. Unterrichtswissenschaft. *Unterrichtswissenschaft, 42*, 280 – 288.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed.). New York, NY: The Guilford Press.

König, J. & Blömeke, S. (2009). Pädagogisches Wissen von angehenden Lehrkräften. Erfassung und Struktur von Ergebnissen der fachübergreifenden Lehrerausbildung. *Zeitschrift für Erziehungswissenschaft, 12*, 499 – 527.

König, J., Ligtvoet, R., Klemenz, S. & Rothland, M. (2017). Effects of Opportunities to Learn in Teacher Preparation on Future Teachers' General Pedagogical Knowledge: Analyzing Program Characteristics and Outcomes. *Studies in Educational Evaluation, 53*, 122 – 133.

Kunina-Habenicht, O., Lohse-Bossenz, H., Kunter, M., Dicke, T., Förster, D., Gößling, J. et al. (2012). Welche bildungswissenschaftlichen Inhalte sind wichtig in der Lehrerbildung? Ergebnisse einer Delphi-Studie. *Zeitschrift für Erziehungswissenschaft, 15*, 649 – 682.

Kunina-Habenicht, O., Maurer, C., Schulze-Stocker, F., Wolf, K., Hein, N., Leutner, D., Seidel, T. & Kunter, M. (2019). Zur curricularen Validität des BilWiss 2.0-Tests zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften. *Zeitschrift für Pädagogik, 65*, 542 – 556.

Kunina-Habenicht, O., Schulze-Stocker, F., Kunter, M., Baumert, J., Leutner, D., Förster, D. et al. (2013). Die Bedeutung der Lerngelegenheiten im Lehramtsstudium und deren individuelle Nutzung für den Aufbau des bildungswissenschaftlichen Wissens. *Zeitschrift für Pädagogik, 59* (1), 1 – 23.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T. & Hachfeld, A. (2013). Professional Competence of Teachers: Effects on Instructional Quality and Student Development. *Journal of Educational Psychology, 105*, 805 – 820.

Kunter, M., Kunina-Habenicht, O., Dicke, T., Holzberger, D., Lohse-Bossenz, H., Leutner, D. et al. (2017). Bildungswissenschaftliches Wissen und professionelle Kompetenz in der Lehramtsausbildung: Ergebnisse des Projekts BilWiss. In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals. Interdisziplinäre Betrachtungen, Befunde und Perspektiven* (S. 37 – 54). Heidelberg: Springer VS.

Lange, K., Ohle, A., Kleickmann, T., Kauertz, A., Möller, K. & Fischer, H. (2015). Zur Bedeutung von Fachwissen und fachdidaktischem Wissen für Lernfortschritte von Grundschülerinnen und Grundschülern im naturwissenschaftlichen Sachunterricht. *Zeitschrift für Grundschulforschung, 8* (1), 23 – 38.

Lenske, G., Thillmann, H., Wirth, J., Dicke, T. & Leutner, D. (2015). Pädagogisch-psychologisches Professionswissen von Lehrkräften: Evaluation des ProwiN-Tests. *Zeitschrift für Erziehungswissenschaft*, 1 – 21.

Linninger, C., Kunina-Habenicht, O., Emmenlauer, S., Dicke, T., Schulze-Stocker, F., Leutner, D. et al. (2015). Assessing Teachers' Educational Knowledge: Construct Specification and Validation Using Mixed Methods. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*, 72 – 83.

Lohse-Bossenz, H., Kunina-Habenicht, O., Dicke, T., Leutner, D. & Kunter, M. (2015). Teachers' knowledge about psychology: Development and validation of a test measuring theoretical foundations for teaching and its relation to instructional behavior. *Studies in Educational Evaluation, 44*, 36 – 49.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*, 159 – 176.

Müser, S., Fleischer, J. & Leutner, D. (2018). *Die KMK-Standards in der Lehrerbildung. Konzeption und Validierung eines Messinstruments* (Vortrag auf der Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung AEPF, Lüneburg, 24. – 26. 09. 2018).

Muthén, L. K. & Muthén, B. (1998 – 2017). *Mplus* (Version 8). Los Angeles, CA: Muthén & Muthén.

Oser, F. (2001). Standards: Kompetenzen von Lehrpersonen. In F. Oser & J. Oelkers (Hrsg.), *Die Wirksamkeit der Lehrerbildungssysteme. Von der Allrounderbildung zur Ausbildung professioneller Standards* (S. 215 – 336). Zürich: Rüegger.

Robitzsch, A., Kiefer, T. & Wu, M. (2017). *TAM: Test analysis modules.* R package (Version 2.6 – 2).

Schleicher, A. (2016). *Teaching Excellence through Professional Learning and Policy Reform.* Paris: OECD.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (5., vollst. überarb., akt. und erw. Aufl.). Heidelberg: Springer.

Schulze-Stocker, F., Holzberger, D., Kunina-Habenicht, O., Terhart, E. & Kunter, M. (2016). Spielen Studienschwerpunkte wirklich eine Rolle? Zum Zusammenhang von bildungswissenschaftlichen Studienschwerpunkten, selbsteingeschätzten Kenntnissen und gemessenem Wissen am Ende eines Lehramtsstudiums. *Zeitschrift für Erziehungswissenschaft, 19*, 599 – 623.

Schulze-Stocker, F., Holzberger, D. & Lohse-Bossenz, H. (2017). Das bildungswissenschaftliche Curriculum – Zentrale Ergebnisse des BilWiss-Programms. *Das Hochschulwesen, 65* (4+5), 134 – 138.

Seifert, A., Hilligus, A. H. & Schaper, N. (2009). Entwicklung und psychometrische Überprüfung eines Messinstruments zur Erfassung pädagogischer Kompetenzen in der universitären Lehrerbildung. *Lehrerbildung auf dem Prüfstand, 2*, 82 – 103.

Seifert, A. & Schaper, N. (2010). Überprüfung eines Kompetenzmodells und Messinstruments zur Strukturierung allgemeiner pädagogischer Kompetenz in der universitären Lehrerbildung. *Lehrerbildung auf dem Prüfstand, 3*, 179 – 198.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15* (2), 4 – 21.

Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). (2014). *Standards für die Lehrerbildung: Bildungswissenschaften. vom 16. 12. 2004 i. d. F. vom 16. 05. 2019.* https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf

Tachtsoglou, S. & König, J. (2017). Der Einfluss universitärer Lerngelegenheiten auf das pädagogische Wissen von Lehramtsstudierenden. *Zeitschrift für Bildungsforschung, 7*, 291 – 310.

Terhart, E. (2012). „Bildungswissenschaften": Verlegenheitslösung, Sammeldisziplin, Kampfbegriff. *Zeitschrift für Pädagogik, 58* (1), 22 – 39.

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54*, 143 – 178.

Voss, T., Kunina-Habenicht, O., Hoehne, V. & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften: Empirische Zugänge und Befunde. *Zeitschrift für Erziehungswissenschaft, 18*, 187 – 223.

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

Xia, Y. & Yang, Y. (2018). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*, 409 – 428. https://doi.org/10.3758/s13428-018-1055-2

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Los Angeles, CA: University of California.

Jun. Prof. Dr. Olga Kunina-Habenicht
Institut für bildungswissenschaftliche Forschungsmethoden
Fakultät für Geistes- und Humanwissenschaften
Pädagogische Hochschule Karlsruhe
Bismarckstraße 10
76133 Karlsruhe
olga.kunina-habenicht@ph-karlsruhe.de

Dr. Christina Maurer
Kristin Wolf
Prof. Dr. Mareike Kunter
Fachbereich 05 Psychologie & Sportwissenschaft
Institut für Psychologie
Goethe-Universität Frankfurt am Main
Theodor-W.-Adorno-Platz 6
60629 Frankfurt a. M.

Prof. Dr. Doris Holzberger
Maria Schmidt
Prof. Dr. Tina Seidel
TUM School of Education
Technische Universität München (TUM)
Arcisstraße 21
80333 München

Dr. Theresa Dicke
Institute for Positive Psychology and Education
Australian Catholic University
PO Box 968
North Sydney NSW 2059, Australia

Ziwen Teuber
Abteilung für Psychologie
Fakultät für Psychologie und Sportwissenschaft
Universität Bielefeld
Postfach 10 01 31
33501 Bielefeld

Dr. Marta Koc-Januchta
Prof. Dr. Detlev Leutner
Fakultät für Bildungswissenschaften
Lehrstuhl für Lehr-Lernpsychologie
Universität Duisburg-Essen
Universitätsstraße 2
45141 Essen

Jun. Prof. Dr. Hendrik Lohse-Bossenz
Pädagogische Hochschule Heidelberg
Keplerstraße 87
69120 Heidelberg

