Blömeke, Sigrid [Hrsg.]; Zlatkin-Troitschanskaia, Olga [Hrsg.]

Kompetenzen von Studierenden

Weinheim u.a. : Beltz Juventa 2015, 235 S. - (Zeitschrift für Pädagogik, Beiheft; 61)

61. Beiheft April 2015

ZEITSCHRIFT FÜR PÄDAGOGIK

Kompetenzen von Studierenden

BELTZ JUVENTA

Zeitschrift für Pädagogik · 61. Beiheft

Zeitschrift für Pädagogik · 61. Beiheft

Kompetenzen von Studierenden

Herausgegeben von
Sigrid Blömeke und Olga Zlatkin-Troitschanskaia

BELTZ JUVENTA

www.beltz.de · www.juventa.de
Herstellung: Lore Amann
Satz: text plus form, Dresden

E-Book
ISSN 0514-2717
Bestell-Nr. 443508

Inhaltsverzeichnis

Forschungsbezogene Kompetenzen

Sigrid Blömeke/Olga Zlatkin-Troitschanskaia

Kompetenzen von Studierenden

Einleitung

Die empirische Bildungsforschung hat in der Vergangenheit vor allem schulisches Lernen und seine Erträge in den Blick genommen. Die jüngsten tiefgreifenden Umstrukturierungen im Hochschulsektor – nicht zuletzt im Kontext der Bologna-Reform, des Europäischen Qualifikationsrahmens oder der Exzellenzinitiative – haben dazu geführt, dass auch die Erforschung akademisch erworbener Kompetenzen in den letzten Jahren an Bedeutung gewonnen hat (vgl. Blömeke, Zlatkin-Troitschanskaia, Kuhn & Fege, 2013). Vor diesem Hintergrund hat das Bundesministerium für Bildung und Forschung (BMBF) eine Förderinitiative „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor" (KoKoHs) eingerichtet, in der 23 Verbundprojekte an über 70 Hochschulstandorten bundesweit seit 2011 zur theoretischen Modellierung, Operationalisierung und Validierung von Testinstrumenten zur Erfassung von Kompetenzen forschen, die im akademischen Kontext erworben werden (vgl. Blömeke & Zlatkin-Troitschanskaia, 2013). Im vorliegenden Beiheft der *Zeitschrift für Pädagogik* werden wichtige Haupterträge dargestellt, indem ausgewählte Projekte ihre empirischen Ergebnisse zu den „Kompetenzen von Studierenden" vor dem Hintergrund eines einheitlichen theoretischen Rahmens darstellen.

Alle Beiträge sind von übergreifender Bedeutung und lassen sich einer der beiden folgenden Perspektiven zuordnen, die typisch für die universitäre Ausbildung sind: die theoretische Modellierung und empirische Erfassung *berufsbezogener* akademischer Kompetenzen sowie die Modellierung und Erfassung *wissenschaftsbezogener* Kompetenzen. Erstere beziehen sich auf Kompetenzen, wie sie typischerweise in anwendungsorientierten Bachelor- und Masterstudiengängen erworben werden (sollten). Letztere decken zum einen Kompetenzen ab, die speziell in forschungsorientierten Studiengängen erworben werden (sollten), zum anderen aber auch solche studienfachübergreifenden Kompetenzen, die generell an einer Hochschule als Ort akademischer Ausbildung vermittelt werden (sollten).

Vier der berufsbezogenen Beiträge decken pädagogische Berufe ab und sind insofern ein Indikator des Aufschwungs empirischer Lehrerbildungsforschung. Hammer et al. haben ein Testinstrument für die Erfassung von Kompetenzen Lehramtsstudierender im Bereich *Deutsch als Zweitsprache* entwickelt, dessen Reichweite sie in ihrem Beitrag „Kompetenz von Lehramtsstudierenden in Deutsch als Zweitsprache: Validierung des GSL-Testinstruments" anhand von Ergebnissen zur Kompetenzstruktur von DaZ und ihrer Relation zu Lerngelegenheiten vorstellen. Riese et al. erfassen die Kompetenzen von angehenden *Physiklehrkräften* unter anderem mithilfe eines innovativen Testformats in Form von praxisnahen Rollenspielen, um Lehrerhandeln in Laborsitua-

tionen abbilden zu können. In ihrem Beitrag „Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik“ berichten sie Ergebnisse zur Struktur dieses Wissens und ihrer Relation zu ähnlichen und unterschiedlichen Konstrukten.

Dunekacke et al. untersuchen in ihrem Beitrag „Mathematikdidaktische Kompetenz von Erzieherinnen und Erziehern: Validierung des KomMa-Leistungstests durch die videogestützte Erhebung von Performanz“ den Zusammenhang von kognitiven Leistungsdispositionen und situationsnahen Fähigkeiten. Unter anderem stellen sie einen neu entwickelten videobasierten Test vor, mit dem sie das Feld der *Frühpädagogik* erschließen. Bouley et al. schließlich untersuchen in ihrem Beitrag „Der Einfluss von universitären und außeruniversitären Lerngelegenheiten auf das Fachwissen und fachdidaktische Wissen von angehenden Lehrkräften an kaufmännisch-berufsbildenden Schulen“ die Kompetenzen von Studierenden der *Wirtschaftspädagogik* mit einem neu entwickelten Testinstrument zum Rechnungswesen. Im Beitrag von Zlatkin-Troitschanskaia et al. wird der Analysefokus auf die Domäne der Wirtschaftswissenschaften ausgeweitet. In ihrem Beitrag „Erwerb wirtschaftswissenschaftlicher Fachkompetenz im Studium – Eine mehrebenenanalytische Betrachtung von hochschulischen und individuellen Einflussfaktoren“ wird mit einem Mixed-Methods-Ansatz der Einfluss hochschulischer Lehrangebote und weiterer institutioneller Merkmale (z. B. Art der Hochschule) auf das *volkswirtschaftswissenschaftliche Fachwissen* der Studierenden fokussiert.

Kompetenzen, die vertieft in forschungsorientierten Studiengängen erworben werden, deren Grundlegung jedoch auch in allen übrigen hochschulischen Studiengängen erfolgen sollte, sind Gegenstand der zweiten Forschungssäule der KoKoHs-Förderinitiative und dieses Beiheftes. Hier geht es beispielsweise darum zu diagnostizieren, inwieweit Studierende Forschungsliteratur *rezipieren* können, inwieweit sie selbst *forschen* können und inwiefern sie in der Lage sind, wissenschaftlich zu *schreiben.*

Trempler, Hetmanek et al. untersuchen die Kompetenz von zukünftigen Lehrpersonen, evidenzbasiert zu argumentieren. In ihrem Beitrag „Nutzung von Evidenz im Bildungsbereich – Validierung eines Instruments zur Erfassung von Kompetenzen der Informationsauswahl und Bewertung von Studien“ stellen sie ihr Diagnoseinstrument und die Operationalisierung der Subdimensionen *Informationsauswahl* und *Bewertung von Studien* vor, für deren Konstrukt-, Inhalts- und Kriteriumsvalidität sie empirische Belege vorlegen. Welche *Forschungskompetenz* die Gruppe an bildungswissenschaftlichen Studierenden – vor allem des Lehramts – besitzt, untersuchen Schladitz et al. in ihrem Beitrag „Konstruktvalidierung eines Tests zur Messung bildungswissenschaftlicher Forschungskompetenz“. Als diskriminanten Validierungszugang grenzen sie diese Kompetenz von fluider Intelligenz ab, als konvergente Validierung erwarten sie mittlere bis hohe Zusammenhänge zur selbsteingeschätzten Forschungskompetenz. Winter-Hölzl et al. entwickeln und validieren schließlich ein Diagnoseinstrument, um zu prüfen, wie vertraut Studierende der Bildungswissenschaften mit dem *wissenschaftlichen Schreiben* sind. In ihrem Beitrag „Entwicklung und Validierung eines Tests zur Erfassung des Genrewissens Studierender und Promovierender der Bildungswissenschaften“ stellen sie vor, welches explizierbare rhetorische Wissen und inwieweit diagnostisches rhetorisches Wissen vorhanden ist. Steuer et al. schließlich entwickeln und validieren

ein Instrument zur Erfassung der Kompetenzen von Studierenden zum selbstregulierten Lernen in unterschiedlichen Studiengängen.

Alle Beiträge nehmen im ersten Teil eine theoriegeleitete Modellierung des zu erfassenden Konstrukts vor und präsentieren die entwickelten Instrumente. Den Beiträgen des vorliegenden Beiheftes liegt ein Kompetenzbegriff zugrunde, der an Weinert (2001) sowie Koeppen, Hartig, Klieme und Leutner (2008) anschlussfähig ist. Kompetenzen werden als *latente Leistungsdispositionen* angesehen, die von Performanz abzugrenzen sind. Es handelt sich um *erlernbare,* nicht um angeborene Fähigkeiten oder gar anthropologische Konstanten. Dies bedeutet zugleich, dass unter Entwicklungsgesichtspunkten von Erweiterungs-, aber auch von Prozessen des Vergessens oder Absinkens ausgegangen werden kann bzw. muss. Kompetenzen werden über die Zeit und Einzelsituationen hinweg als relativ stabile *(trait)* Dispositionen betrachtet, die von dynamischen *(state)* Komponenten beeinflusst werden können.

Allen Beiträgen liegt zudem die Annahme zugrunde, dass die akademisch erworbenen Kompetenzen *mehrdimensional* und – trotz generischer Anteile wie z. B. forschungsmethodische Fähigkeiten – *domänenspezifisch* ausgeprägt sind und sich damit (auch in den generischen Anteilen) von Konstrukten wie Intelligenz bzw. kognitiven Grundfähigkeiten unterscheiden lassen. Die Kompetenzmodellierung nimmt ihren Ausgangspunkt einerseits bei den an den Hochschulen vermittelten Kerninhalten, andererseits bei typischen beruflichen Anforderungssituationen, die idealerweise im Einklang stehen.

Mit diesem mehrperspektivischen Herangehen an die Kompetenzmodellierung im Hochschulsektor wird zugleich eine Herausforderung der Projekte im Vergleich zur Kompetenzmodellierung im Schulsektor angesprochen. In den meisten Studiendomänen existieren keine bundesweiten „Kerncurricula". Zudem ist i. d. R. von einem Spannungsverhältnis zwischen einer curricularen (aus der Sicht der akademischen Disziplinen geprägten) und einer berufsbezogenen Perspektive auszugehen. Die Modelle sind angesichts der hohen institutionellen und curricularen Heterogenität des deutschen Hochschulwesens sowie der Variabilität und Dynamik der Arbeitsmärkte vermutlich *nicht zwingend* für jeden einzelnen Studiengang und jeden einzelnen Arbeitsplatz valide, sondern bilden zentrale Inhalte und Anforderungen der jeweils fokussierten Domäne ab.

Alle Beiträge folgen einer mehrdimensionalen Kompetenzorientierung, legen den Fokus jedoch i. d. R. auf die Modellierung *kognitiver* Kompetenzen von Studierenden sowie die *Konstruktvalidierung.* Diese theoretisch und methodisch bedeutsame Frage, ob für die Kompetenzmessung im Hochschulsektor entwickelte Instrumente tatsächlich das erfassen, was von ihnen erwartet wird und worauf sich die spätere Verwendung und Interpretation der gewonnenen Testscores gründen würde, schien uns im Hinblick auf das komplexe Feld der Hochschulforschung von weitreichender Bedeutung.

Die beiden Stränge im vorliegenden Beiheft werden jeweils durch einen Kommentar begleitet, in dem eine projektübergreifende zusammenfassende Betrachtung und kritische Reflexion der erzielten Ergebnisse erfolgt und Konsequenzen für die weitere Forschung herausgearbeitet werden. Der Kommentar von Kaiser bezieht sich auf den fach-

bzw. berufsbezogenen Strang des Beiheftes, derjenige von König auf den generischen bzw. forschungsbezogenen Strang.

Da Validität das fundamentale und zugleich komplexeste Gütekriterium empirischer Bildungsforschung ist, stellt ihr Nachweis den gemeinsamen Anspruch aller Beiträge in diesem Beiheft dar. Drei bis heute gebräuchliche Validitätskriterien weisen die längste Tradition in der Leistungsdiagnostik auf (Frey, 2013): die Kriteriumsvalidierung, die Inhaltsvalidierung und die Konstruktvalidierung von Testverfahren. Konstruktvalidität umfasst die empirischen Befunde und Argumente, mit denen die Zuverlässigkeit der Interpretation von Testergebnissen im Sinne erklärender Konzepte gestützt wird, die sowohl die Testergebnisse selbst als auch die Zusammenhänge der Testwerte mit anderen Variablen erklären (Messick, 1995). In ihrem einleitenden Beitrag fassen Jenßen, Dunekacke und Blömeke den aktuellen Diskussions- und Forschungsstand zur Validierung als einem zentralen Element der „Qualitätssicherung in der Kompetenzforschung" zusammen. Sie lassen ihre Erkenntnisse in „Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis" münden, auf die in den nachfolgenden Beiträgen Bezug genommen wird.

Literatur

Blömeke, S., & Zlatkin-Troitschanskaia, O. (Hrsg.) (2013). *The German funding initiative „Modeling and Measuring Competencies in Higher Education: 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students"* (KoKoHs Working Papers, 3). Berlin/Mainz.

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., & Fege, J. (Hrsg.) (2013). *Modeling and Measuring Competencies in Higher Education*. Rotterdam: Sense Publishers.

Frey, A. (2013). *Validität.* Eröffnungsvortrag im Rahmen des KoKoHs-Rundgesprächs zu Validität und Validierung am 14. März 2013 an der Humboldt-Universität zu Berlin.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Zeitschrift für Psychologie, 216*, 61–73.

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

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45–66). Göttingen: Hogrefe.

Anschrift der Autorinnen

Prof. Dr. Sigrid Blömeke, Centre for Educational Measurement at the University of Oslo (CEMO), Leibniz-Institut für die Pädagogik der Mathematik und Naturwissenschaften Kiel, Humboldt-Universität zu Berlin, Postboks 1072/Blindern, 0316 Oslo, Norwegen
E-Mail: sigribl@cemo.uio.no

Prof. Dr. Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-Universität Mainz, Fachbereich Rechts- und Wirtschaftswissenschaften, Jakob Welder-Weg 9, 55099 Mainz, Deutschland
E-Mail: lstroitschanskaia@uni-mainz.de

Lars Jenßen/Simone Dunekacke/Sigrid Blömeke

Qualitätssicherung in der Kompetenzforschung

Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis

Zusammenfassung: Der Beitrag diskutiert Anforderungen an die Validität von Messverfahren zur Erfassung von Kompetenzen und fokussiert speziell den Nachweis von Inhaltsvalidität. Nach einer theoretischen Einbettung dieser in das Spektrum an Validierungsnotwendigkeiten wird exemplarisch anhand des Projekts *KomMa* ein Verfahren zur Inhaltsvalidierung eines Kompetenztests vorgestellt. Mithilfe externer Expertinnen und Experten aus Forschung und Praxis wurde ein ökonomisches schriftliches Ratingverfahren durchgeführt, um die inhaltliche Qualität der konstruierten Items sicherzustellen. Das im Projekt *KomMa* entwickelte Verfahren zur Inhaltsvalidierung wird Schritt für Schritt skizziert und zur Diskussion gestellt. Praktische Hinweise für die Durchführung einer Inhaltsvalidierung und Empfehlungen für verschiedene Validierungsstrategien gegliedert nach Validitätsaspekten schließen den Beitrag ab.

Schlagworte: Validität, Test, Inhaltsvalidierung, Expertenrating, Kompetenzmessung

1. Einleitung

Dass die in empirischen Studien gewonnenen Ergebnisse zum Beispiel zur professionellen Kompetenz von Lehrkräften Gütekriterien erfüllen müssen, gehört zum Lehrbuchwissen in der Bildungsforschung (Bortz & Döring, 2006; Rost, 2004). Die Gültigkeit (Validität) ist dabei das fundamentale und zugleich komplexeste Gütekriterium. Erstaunlich wenig elaboriert sind dabei jedoch insbesondere Hinweise, wie Inhaltsvalidität als eine unverzichtbare Voraussetzung für andere Validitätsfacetten nachgewiesen werden kann, wenngleich ihre Bedeutung gerade für die Testentwicklung seit langer Zeit betont wird (Messick, 1989).

Der vorliegende Beitrag zielt vor diesem Hintergrund zum einen darauf, die vorhandenen Standards für die Validierung zusammenzufassen und so eine konzeptionelle Rahmung des Themas „Validität in der Kompetenzmessung“ zu leisten. Zum anderen zielt der Beitrag darauf, konkrete Verfahren zur Sicherung von Inhaltsvalidität vorzustellen und diese für die Testentwicklung bzw. die Veröffentlichungspraxis zu empfehlen sowie beispielhaft empirisch zu illustrieren. Eine Sichtung der Veröffentlichungen der letzten zehn Jahre hat deutlich gemacht, dass bei Kompetenztests in der Regel zwar ausführlich der konzeptionelle Rahmen und das Item-Framework beschrieben werden (z. B. Blömeke, Kaiser & Lehmann, 2008; Brunner et al., 2006). Ob die inhaltliche Überführung des theoretischen Rahmens in Testaufgaben systematisch validiert wurde, bleibt allerdings oft unklar. Mit wenigen Ausnahmen (z. B. Lohse-Bossenz, Kunina-

Habenicht & Kunter, 2013; Watermann & Klieme, 2002) sind die Hinweise nur allgemein und lassen vermuten, dass dieser Schritt eher rudimentär stattgefunden hat, indem lediglich die Testentwickler selbst, ggf. unter Herbeiziehung ihres engen Umfeldes, eine Inhaltsvalidierung durchgeführt haben.

Aus diesem Defizit kann nicht nur eine Reihe methodischer Probleme resultieren, sondern diese Lücke führt zum einen zu ungenügender Anschlussfähigkeit der verschiedenen Studien untereinander sowie zum anderen in der breiten *scientific community* zu andauernden Zweifeln daran, ob die entwickelten Kompetenztests tatsächlich das erfassen, was sie erfassen sollen (z. B. Rindermann, 2006). Dabei hängt die Akzeptanz eines konstruierten Kompetenztests wesentlich von dessen konzeptioneller Überzeugungskraft ab.

2. Theoretischer Hintergrund

2.1 Validierung von Testverfahren[1]

Das Verständnis, was unter Validierung zu fassen ist, hat sich in den letzten Jahrzehnten stark gewandelt. Die Empfehlungen der Fachverbände American Educational Research Association (AERA), American Psychological Association (APA) und National Council on Measurement in Education (NCME) zu Teststandards spiegeln diesen Wandel wider (Frey, 2013). Wurde in der ersten Auflage (APA, 1954) Validität noch als Eigenschaft eines Tests angesehen, rückt mittlerweile in der vierten Auflage der „Standards for Educational and Psychological Testing" (AERA, APA & NCME, 1985) stärker die Testwert*interpretation* in den Vordergrund. Validierung wird demnach definiert als „the approriateness, meaningfulness, and usefulness of specific inferences made from test scores" (S. 9). Inhalts-, Kriteriums- und Konstruktvalidität stellen hierfür verschiedene Arten an Evidenz dar (Frey, 2013). Kane (1992) lenkte den Blick weg von der alleinigen Absicherung durch formalisierte Theorien stärker hin zu einer überzeugenden, theoretisch fundierten argumentativen Stützung der Testwertinterpretation *(argument-based approach to validation).* Validität ist danach „an integrated evaluation judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (S. 13).

Dieses Validitätsverständnis wurde von Kane (2013) im Anschluss an Arbeiten von Messick und Cronbach später nochmals erweitert, indem neben die messtheoretische Perspektive mit dem Nachweis der Präzision einer Messung, ihrer Konstruktvalidität und der angemessenen Interpretation von Testscores eine ethische Perspektive trat. In dieser werden Fragen nach Kosten und Nutzen von Testungen sowie nach absichtli-

1 Die Abschnitte 2.1, 3.1 und 3.2 stellen eine stark gekürzte und überarbeitete Fassung von Blömeke (2013) dar.

chen und unabsichtlichen Folgen bzw. Nebenwirkungen im Sinne einer *policy perspective* aufgeworfen. Messick (1989) hatte diese Form der Validierung als *consequential validity* bezeichnet. Diese Perspektive ist heute die umstrittenste – nicht nur, weil sie schwierig empirisch zu prüfen ist. Sie macht die Testentwickler auch verantwortlich für spätere Einsätze ihrer Tests, ggf. auch zu gänzlich anderen Zwecken als ursprünglich geplant.

Was Kane (2013) mit dieser Erweiterung des Validitätsbegriffs erreichen will, ist – an jene gerichtet, die einen Test für einen neuen Zweck einsetzen wollen – die Verpflichtung, die ursprünglichen Testentwickler bei diesen neuen Verwendungen einzubeziehen. Die Testentwickler hätten die größte Expertise, um zu beurteilen, inwieweit solche Ausweitungen angemessen und gerechtfertigt seien. Kane (2013, S. 52) verdeutlicht in diesem Zusammenhang zudem, dass „Testing programs have long been known to have strong effects on how schools function, on how and what teachers teach, and on what students study and learn".

Als eigener Aspekt sei in diesem Zusammenhang darauf verwiesen, dass für die Abschätzung der Konsequenzen von Testverfahren *(consequential validity)* auch berücksichtigt werden muss, ob der Test für die Individualdiagnostik oder zu Forschungs- bzw. Systemevaluationszwecken auf Gruppenebene zum Beispiel im Rahmen von Large-Scale-Assessments konstruiert wird. Während im letzten Fall vor allem systembezogene Konsequenzen im Sinne nicht-intendierter Nebenwirkungen diskutiert werden, muss im ersten Fall gegebenenfalls von bedeutsamen Konsequenzen für das Individuum ausgegangen werden, beispielsweise im Rahmen von Selektionsprozessen in der Eignungsdiagnostik (Wottawa & Hossiep, 1997).

Gegen dieses umfassende Verständnis von Validität wird allerdings auch Widerspruch erhoben. Borsboom, Mellenbergh und van Heerden (2004) oder Scriven (2010) betonen, dass ein engeres Konzept notwendig sei, um den Begriff nicht zu überladen und ihm damit seine Eindeutigkeit zu nehmen. Borsboom et al. (2004) führen zudem wissenschaftstheoretische Argumente an, warum sie die Validierung vor allem der Instrumente selbst für bedeutsam halten. Die von einem Test erfasste Eigenschaft existiere unabhängig vom Testverfahren, und ein Test sei dann valide, „if variation in the attribute causes variation in the test scores" (S. 1067). Sie proklamieren also zum einen eine beobachtungsunabhängige Existenz von Eigenschaften und zum anderen eine explizit kausale Beziehung zwischen Konstruktausprägung und Testwert.

Zusammenfassend kann festgehalten werden, dass es im Zusammenhang der Kompetenzmodellierung und Kompetenzerfassung dem Stand der Forschung entspricht, die Validität verschiedener *Interpretationen* von Testergebnissen nachzuweisen, statt nur von „der Validität eines Tests" zu sprechen. Im ersten Validierungsschritt gilt es daher zu spezifizieren, auf welche Interpretation eines Testergebnisses sich eine Validierung bezieht (Hartig, 2013). Verschiedene Interpretationen können sich zum Beispiel auf die Punktevergabe in einem Test (z. B. die Rechtfertigung von Auswertungsschlüsseln und Durchführungsprozeduren), das Verallgemeinern des Ergebnisses (auf nicht im Test enthaltene, aber ähnliche Aufgaben unter ähnlichen Bedingungen), das Extrapolieren über das Testergebnis hinaus (auf andere Kontexte und Aufgabenformate), das kausale Er-

klären eines Testwertes und auf das Treffen weiterführender Entscheidungen als Konsequenz aus dem Testergebnis beziehen (Kane, 2001).

2.2 Das Konzept der Inhaltsvalidität

Das Verständnis von Inhaltsvalidität hat sich im Zuge der „Metamorphose" des Validitätsbegriffs ebenfalls verändert (Geisinger, 1992). Inhaltsvalidität stellt heute (AERA, APA & NCME, 1999) eine Art der Evidenz dar, die neben den übrigen Validitätsarten dazu dient, die Gültigkeit intendierter Testwertinterpretationen zu stützen. Sie ist dabei der Konstruktvalidität untergeordnet (Guion, 1977) und befasst sich mit der Frage, „inwieweit die Inhalte eines Tests bzw. der Items, aus denen er sich zusammensetzt, tatsächlich das interessierende Merkmal erfassen" (Hartig, Frey & Jude, 2012, S. 148). Die Inhaltsvalidität hat, je nach Über- oder Unterrepräsentanz der Inhalte eines Konstrukts im Test bzw. Item, somit direkten Einfluss auf die Konstruktvalidität (Kane, 2013), da Items, denen keine Inhaltsvalidität bescheinigt werden kann, zu unbrauchbaren Ergebnissen führen (Rossiter, 2008).

Das Ergebnis einer Inhaltsvalidierung beruht auf fachlichen Überlegungen und besteht – im Gegensatz zu kriterialen Validierungen oder Konstruktvalidierungen – aus subjektiven Einschätzungen (Bortz & Döring, 2006, S. 200). Nach Klauer (1984) ist die Frage, inwieweit die Items bzw. der Test die Grundgesamtheit des Konstrukts abbilden, dabei handlungsleitend bei der Einschätzung. Hartig, Frey und Jude (2012) verwenden hierfür den Begriff des Repräsentationsschlusses als Ziel der Inhaltsvalidierung. Die Methode der Wahl für die Inhaltsvalidität besteht folglich in der systematischen Befragung von „Experten" – wobei ein Nachweis für die Expertise anhand geeigneter Kriterien zu erbringen ist (Hornke & Winterfeld, 2004).

Welche Fragen an die Expertinnen und Experten im Zuge der Inhaltsvalidierung gestellt werden müssen, orientiert sich an der Art der Definition des zu messenden Konstrukts, ob dieses operational oder theoretisch definiert wird (Hartig, Frey & Jude, 2012). Wird das Konstrukt operational definiert, stellen die Iteminhalte die Definition des Konstrukts dar („der Test misst, was er misst"). Bei der vorzuziehenden theoretisch begründeten Definition liegen Vorstellungen zur Struktur und zum Inhalt des Konstrukts vor. Hypothesen über Merkmale des Konstrukts, auf die Unterschiede in seiner Varianz zurückzuführen sind, werden theoriegeleitet begründet. Bei operational definierten Konstrukten zielt die Inhaltsvalidierung vor allem auf Verallgemeinerungsschlüsse von Testergebnissen auf eine Domäne ab, während bei theoretisch definierten Konstrukten erklärende Aussagen im Fokus stehen (Kane, 2001). In der Testpraxis stellen die Ansätze Pole auf einem Kontinuum dar, weil es meist an umfassenden Theorien über Personenunterschiede in einem bestimmten Merkmal mangelt (Hartig, Frey & Jude, 2012).

Eine inhaltliche Validierung kann auf die Test- und auf die Itemebene bezogen werden, beispielsweise auf die Angemessenheit der Operationalisierung allgemeiner mathematischer Kompetenz durch den einzusetzenden Itempool als Ganzen (Testebene) oder

auf die konkret umgesetzte Kombination von „Problemlösen“ und „Zahlen, Mengen und Operationen“ in einem Item (Itemebene). Auf Itemebene stellen sich nach Hartig, Frey und Jude (2012) die Fragen, inwieweit das Item Teil des interessierenden Itemuniversums ist (z. B. alle theoretisch möglichen Items einer Kombination von „Problemlösen“ und „Zahlen, Mengen und Operationen“) und in welchem Ausmaß es als prototypisch für diese Gesamtheit angesehen werden kann. Des Weiteren muss geklärt werden, ob das Item den intendierten Inhalt tatsächlich repräsentiert. Dabei sind auch der Itemstamm und das Antwortformat einzubeziehen (Messick, 1989).

2.3 Inhaltsvalidität bei Kompetenztests

Inhaltsvalidität spielt vor allem im Rahmen der Test*entwicklung* eine Rolle (Bortz & Döring, 2006, S. 200). Entscheidend ist dabei, *wofür* ein Test konstruiert wird und *was* mit dem Test untersucht werden soll (Kane, 2013). Kompetenztests können verschiedene Ziele verfolgen: z. B. inwieweit vorgegebene Lehrpläne oder Bildungsstandards erreicht werden, inwieweit Kompetenzentwicklung in einem bestimmten Bildungsabschnitt stattfindet oder inwieweit berufliche Anforderungen in der Praxis bewältigt werden können (Blömeke & Zlatkin-Troitschanskaia, 2013).

Werden Iteminhalte beispielsweise aus Lehrplänen abgeleitet, um die zu erfassende Kompetenz operational zu definieren, steht insbesondere die Frage im Vordergrund, inwieweit ein Item ein dort gefordertes Lernziel erfasst und als wie prototypisch es für einen Lernbereich angesehen werden kann. Kann einem Item in diesem Sinne Inhaltsvalidität bescheinigt werden, dürfte es keine Rolle spielen, wenn in einem Test ein anderes Item aus derselben Domäne verwendet wird. Die Übereinstimmung von Zielen und Inhalten eines Lehrplans mit den Iteminhalten wird als *curriculare Validität* bezeichnet und beschreibt eine Variante der Inhaltsvalidität (Yalow & Popham, 1983). Geht es um diese Variante der Inhaltsvalidierung, muss dies bei der Auswahl der Experten berücksichtigt werden.

Soll der Kompetenztest dafür eingesetzt werden, die Performanz in einer Berufssituation vorherzusagen, besteht das Ziel in einer inhaltlich hohen *trivialen Validität.* Diese Form der Inhaltsvalidität spielt insbesondere bei Eignungsbeurteilungen in beruflichen Kontexten eine Rolle (Lawshe, 1975; Kubinger, 2006, S. 51). Typische berufliche Aufgaben (z. B. Beantwortung von E-Mails) stimmen dementsprechend mit den Inhalten der Diagnostik (z. B. Beantwortung fiktiver E-Mails im Rahmen eines Assessment Centers) überein. Diese Form der Validität kann auch bei Kompetenztests eine Rolle spielen, bei denen mit videogestützten Fallvignetten gearbeitet wird (z. B. Blömeke et al., 2011; Kersting, 2008; Pauli & Reusser, 2006).

Da Kompetenztests für unterschiedliche Ziele eingesetzt werden können, diese in der Phase der Testkonstruktion aber nicht vollends abzusehen sind, sollten verschiedene Formen der Inhaltsvalidierung genutzt werden. Ein Beispiel dafür stellt PISA 2000 dar (Baumert et al., 2003): Das Vorliegen curricularer Validität war in Deutschland zunächst kein Ziel. Erst durch die öffentliche Diskussion zur Übereinstimmung der Testinhalte

mit den Lehrplänen der Bundesländer wurde ein solcher Nachweis nötig und durchgeführt (ebd.).

3. Methoden der Validierung von Testverfahren

3.1 Nachweis von Kriteriumsvalidität

Mit *Kriteriumsvalidität* wird die Vorhersage einer direkt beobachtbaren Verhaltensweise außerhalb der Testsituation als Kriterium für die Gültigkeit eines Diagnoseverfahrens bezeichnet (Schaper, 2013). Je nach Verhaltensdomäne bzw. Konstrukt können unterschiedliche Arten an Kriterien herangezogen werden, und zwar Ergebniskriterien (z. B. Schulnoten oder die Anzahl Vertragsabschlüsse im Versicherungsgeschäft), Verhaltenskriterien (z. B. Ausmaß und Art des Rückmeldeverhaltens von Lehrkräften oder Art und Qualität des kundenorientierten Verhaltens von Servicekräften) und Eigenschaftskriterien (z. B. die Arbeitsmotivation oder das Arbeitsengagement von Mitarbeitern). Ein Beispiel für eine Validierungsstudie im Hochschulsektor stellt die differenzielle Vorhersage von Studienerfolg durch Schulnoten in Abhängigkeit von ihrer Erfassung über Selbstberichte oder eine offizielle Mitteilung durch die Schule dar (Zwick & Himelfarb, 2011).

Zeitlich gesehen kann bei einer kriterialen Validierung zwischen konkurrenter und prognostischer Validität unterschieden werden (Schaper, 2013). Im Falle der Feststellung von konkurrenter oder Übereinstimmungsvalidität wird geprüft, inwieweit die Testwerte mit dem zeitgleich erhobenen Außenkriterium zusammenhängen. Als Beispiel kann der Zusammenhang zwischen einem Test zur sozialen Kompetenz und einer Beurteilung der sozialen Fähigkeit in bestimmten Kontexten durch andere Personen angeführt werden. Im Falle der prognostischen Validität wird geprüft, inwieweit anhand von Testwerten später erhobenes Verhalten oder spätere Leistungen vorhergesagt werden können. Ein Beispiel stellt die Vorhersage von Studienerfolg anhand eines Studieneingangstests dar. Von besonderem Interesse ist dabei, inwieweit das Testverfahren hilfreich ist, wenn es zusätzlich zu bekannten Maßen – zum Beispiel zur Abiturnote, die vergleichsweise leicht erhoben werden kann und deren prognostische Validität für Studienerfolg vielfach belegt ist – eingesetzt wird (inkrementelle Validität).

Liegt kriteriale Validität vor, können also nicht nur Aussagen über die Gültigkeit der mit einem Testverfahren gewonnenen Ergebnisse gemacht werden, sondern es ist auch möglich, Prognosen oder Diagnosen in Bezug auf das Verhalten oder die Leistungsfähigkeit in zukünftigen Kontexten zu machen (Schaper, 2013).

3.2 Der Nachweis von Konstruktvalidität

Konstruktvalidität umfasst die empirischen Befunde und Argumente, mit denen die Zuverlässigkeit der Interpretation von Testergebnissen im Sinne erklärender Konzepte gestützt wird, die sowohl die Testergebnisse selbst als auch die Zusammenhänge der Testwerte mit anderen Variablen erklären (Messick, 1995, S. 743). Drei empirische Strategien zur Stützung von Konstruktvalidität haben sich durchgesetzt (Hartig, 2013): die Prüfung der theoretisch angenommenen inneren Struktur eines Konstrukts *(factorial validity),* die Prüfung der Konstruktrepräsentation über die Vorhersage von Itemschwierigkeiten und die Prüfung der Verortung des Konstrukts in einem nomologischen Netzwerk (einschl. konvergenter und diskriminanter Validität; Campbell & Fiske, 1959).

Die Grundidee einer Prüfung innerer Strukturen ist, dass mit der Entwicklung eines Testinstruments Annahmen über die Dimensionalität des zu erfassenden Konstrukts verbunden sind, die als Hypothesen empirisch überprüft werden können (Hartig, 2013). Annahmen über die Ein- oder Mehrdimensionalität können in Modellen mit latenten Variablen geprüft werden, in denen die angenommenen Strukturen spezifiziert werden, beispielsweise als Strukturgleichungsmodelle oder mehrdimensionale IRT-Modelle.

Die Grundidee der Vorhersage von Itemschwierigkeiten ist, dass mit der Entwicklung eines Testinstruments Annahmen dazu bestehen, welche Anforderungen von Personen mit niedriger, mittlerer oder hoher Kompetenz bewältigt werden können, warum also welche Aufgaben wie schwer sind (z. B. aufgrund kognitiver Prozesse etc.; Embretson, 1983; Hartig & Frey, 2012). Eine empirische Prüfung dieser Annahmen erfolgt, indem Hypothesen darüber formuliert werden, welche Aufgabencharakteristika höhere Anforderungen stellen. Die empirische Aufgabenschwierigkeit wird dann durch die angenommenen Faktoren zu erklären versucht, z. B. in Regressionsanalysen oder erklärenden IRT-Modellen (Hartig, Frey, Nold & Klieme, 2012). Lassen sich die Aufgabenschwierigkeiten (teilweise) erklären, unterstützt dies die Annahme, dass sich die im Test erfassten Kompetenzen durch die Aufgabenanforderungen erklären lassen.

Für die Überprüfung, inwieweit sich das zu erfassende Konstrukt in ein nomologisches Netzwerk einfügen lässt, werden Annahmen darüber formuliert, mit welchen anderen Variablen das zu erfassende Konstrukt in welchem Zusammenhang stehen sollte (Cronbach & Meehl, 1955). Diese Annahmen sind theoriegeleitet zu begründen. Die empirische Prüfung erfolgt, indem die Zusammenhänge des Testwertes mit anderen Variablen zum Beispiel manifest in Form von Korrelationsanalysen oder auf latenter Ebene in Form von Strukturgleichungsmodellen bzw. IRT-Modellen untersucht werden (siehe zum Nachweis von konvergenter und diskriminanter Validität im Rahmen der Multitrait-Multimethod-Methode auch Campbell & Fiske, 1959). Entspricht das Zusammenhangsmuster den theoretisch erwarteten Zusammenhängen, unterstützt dies sowohl die Interpretation der Testwerte bezogen auf das Konstrukt als auch die bei der Spezifikation des nomologischen Netzes herangezogenen theoretischen Annahmen.

Verschiedene Strategien der Konstruktvalidierung schließen sich nicht gegenseitig aus, sondern sollten sich ergänzen. Welche Strategien sich zur Unterstützung der theoriebasierten Interpretation von spezifischen Testwerten empfehlen, hängt davon ab,

wozu präzise theoretische Annahmen existieren, ob also fundierte Hypothesen über eine dimensionale Struktur formuliert werden können oder über Anforderungen, mit denen die Schwierigkeiten von Aufgaben erklärt werden können, bzw. über Zusammenhänge mit anderen Variablen. In neuen Forschungsfeldern wie der Kompetenzerfassung im Hochschulsektor sind die zu untersuchenden Konstrukte mangels empirischer Studien häufig nicht fundiert genug, sodass erste Arbeiten insbesondere in eher unstrukturierten und wenig einheitlich definierten Domänen wie den Geistes- und Sozialwissenschaften notgedrungen eher explorativen Charakter haben (siehe z. B. Blömeke et al., 2011) und echte Validitätsprüfungen noch nicht stattfinden können (Hartig, 2013).

3.3 Der Nachweis von Inhaltsvalidität durch Expertenbefragungen und Expertenpanels

Die bisher dargestellten Verfahren der Kriteriums- und Konstruktvalidierung sind weitgehend etabliert. Evidenz für eine inhaltliche Validität der Messinstrumente liefern sie jedoch noch nicht. Gärtner und Pant (2011) weisen daher auf die oben erläuterten Aspekte der Inhaltsvalidität hin und zeigen auf, dass auch geklärt werden muss, inwieweit gewählte Indikatoren als inhaltlich repräsentativ für das Konstrukt gelten. Sie leisten dies in ihrem Beitrag beispielhaft für den Kontext der Schulinspektionen und schlagen Indikatoren zur Operationalisierung des Konstrukts Schulqualität vor. Dabei betonen sie die Fragen nach rationaler Grundlage, Relevanz und Repräsentanz einzelner Indikatoren als notwendige Bedingungen.

Das Mittel der Wahl bei der Sicherung von Inhaltsvalidität stellt die Begutachtung durch Expertinnen und Experten dar (Popham, 1993; Angoff, 1988, S. 22). Dies kann zum Beispiel durch eine „Delphi-Befragung“ realisiert werden, bei der alle Experten miteinander diskutieren, ob ein Item geeignet ist (Kunina-Habenicht et al., 2012). Diese Methode ist allerdings aufwendig. Expertinnen und Experten können daher auch getrennt befragt und ihre Übereinstimmung auf empirischem Weg festgestellt werden (z. B. Wirtz & Caspar, 2002). Bei der Durchführung von Expertenbefragungen sollte auf typische Schwierigkeiten, wie beispielsweise Beurteilerfehler (Kane, 2013), geachtet werden. In Abschnitt 4 werden exemplarisch derartige Probleme beschrieben und es wird ein möglicher Umgang mit diesen Problemen empfohlen.

Eine systematische Expertenbefragung wurde auch in KomMa durchgeführt. Zur Illustration eines möglichen Verfahrens der Inhaltsvalidierung wird in den folgenden Abschnitten im Anschluss an eine Darlegung des theoretischen Rahmens für das Projekt skizziert, warum die konstruierten Items als inhaltlich valide betrachtet werden können. Dabei spielen für die intendierten Testwertinterpretationen die curriculare Validität und die Validität hinsichtlich zu bewältigender beruflicher Anforderungen eine zentrale Rolle.

4. Nachweis der Inhaltsvalidität im Projekt KomMa

4.1 Modellierung der Kompetenz von Erzieherinnen im Bereich Mathematik

Mit dem gestiegenen gesellschaftlichen und politischen Interesse an frühkindlicher Bildung ist nach der Lehrerausbildung auch die Ausbildung von frühpädagogischen Fachkräften in den Fokus der Bildungsforschung gerückt. International und national sind in diesem Zusammenhang vor allem im Bereich früher mathematischer Bildung Forschungsdesiderate festzustellen (Fried & Roux, 2009; National Mathematics Advisory Panel, 2008).

Bedingt durch das Forschungsdefizit liegen bislang nur wenige und unspezifische Entwürfe für die Modellierung frühpädagogischer professioneller Kompetenz vor. In einem ersten Schritt wird daher der Diskussion in der Lehrerbildungsforschung gefolgt (Speck-Hamdan, 2011). Nach Shulman (1986) kann professionelle Kompetenz in einem Unterrichtsfach als ein Zusammenspiel von Fachwissen, fachdidaktischem und allgemein-pädagogischem Wissen sowie Überzeugungen gesehen werden. Diese Struktur konnte in aktuellen Studien zur deutschen Lehrerausbildung bestätigt werden (Blömeke, Kaiser & Lehmann, 2010). Für die inhaltliche Ausgestaltung dieser Struktur in Bezug auf Erzieherinnen wurden dann in einem zweiten Schritt die pädagogisch-didaktischen Spezifika früher mathematischer Bildung berücksichtigt, die berufliche Anforderungen der zukünftigen pädagogischen Fachkräfte abbilden, ergänzt um aktuelle wissenschaftliche Erkenntnisse. Somit liegt dem KomMa-Kompetenzmodell eine operationale Definition zugrunde.

Die Operationalisierung der beruflichen Anforderungen erfolgte anhand einer halbstandardisierten Analyse der Bildungspläne aller 16 Bundesländer für Kindertageseinrichtungen. Ergebnis dieser Dokumentenanalyse war ein komplexes System beruflicher Anforderungen (Dunekacke et al., 2013). Die Operationalisierung der Wissensfacetten, über die Erzieherinnen verfügen müssen, wenn sie diese Anforderungen bewältigen wollen, erfolgte über eine weitere halbstandardisierte Analyse der Lehrpläne und Studienordnungen aller Ausbildungseinrichtungen für frühpädagogische Fachkräfte in den 16 Bundesländern. Abbildung 1 zeigt das auf dieser Basis entwickelte Kompetenzstrukturmodell.

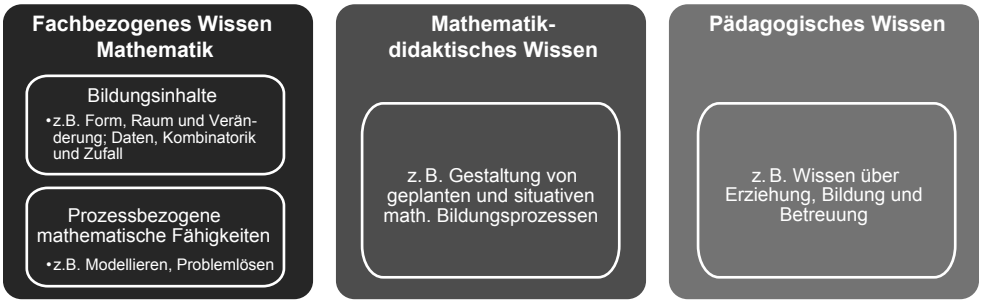


Abb. 1: Kurzfassung der kognitiven Facetten des Kompetenzstrukturmodells in KomMa

4.2 Testentwicklung zur Erfassung der professionellen Kompetenz von Erzieherinnen im Bereich Mathematik

Ziel des Projekts KomMa ist neben der Modellierung der professionellen Kompetenz von frühpädagogischen Fachkräften die Entwicklung eines Leistungstests, um Strukturen dieser Kompetenz untersuchen zu können und Zusammenhänge zu Bedingungsfaktoren (z. B. unterschiedlichen Ausbildungsbedingungen) zu ermitteln. Der Kompetenztest wurde also ausdrücklich nicht für Zwecke der Individualdiagnostik konstruiert.

Da nicht auf vorhandene Erhebungsinstrumente zurückgegriffen werden konnte, war die Neukonstruktion von Testitems erforderlich. Grundlage für die Itemkonstruktion für den Kompetenztest bildete das in Abbildung 1 dargestellte Kompetenzmodell. Die Subdimensionen wurden weiter ausdifferenziert, sodass die Kerninhalte der Items festgelegt waren und aus Sicht der Entwickler das Konstrukt umfassend abbildeten. Die Items sollten zudem ein breites Schwierigkeitsspektrum abbilden.

Zielgruppe des Tests sind angehende frühpädagogische Fachkräfte am Ende der Ausbildung. Hierzu zählen angehende Erzieherinnen, die an Fachschulen ausgebildet werden, sowie angehende Kindheitspädagoginnen, die Bachelorstudiengänge an Fachhochschulen absolvieren. Insgesamt handelt es sich um eine sehr heterogene Zielgruppe (Roth, 2013), deren Heterogenität neben verschiedenen Eingangsvoraussetzungen zusätzlich durch stark differierende Lerngelegenheiten in der Ausbildung erhöht wird.

Aus der Vielzahl an Aufgabenformaten, die für Kompetenztests zur Verfügung steht (Jonkisz, Moosbrugger & Brandt, 2012), wurde in KomMa aus forschungsökonomischen Gründen auf halboffene und geschlossene Formate zurückgegriffen (Bortz & Döring, 2006). Halboffene Formate erfordern das eigenständige Antworten der Testteilnehmenden, wobei die Fragen so formuliert sind, dass anhand von Kodieranweisungen Bewertungen als *richtig* oder *falsch* vorgenommen werden können. Darüber hinaus wurden Multiple-Choice-Items verwendet, bei denen die Teilnehmenden aus mehreren vorgegebenen Antworten die richtige auswählen mussten. Um eine Bearbeitung nach dem Ausschluss-Prinzip möglichst zu vermeiden, wurde der Auswahl geeigneter Distraktoren hinsichtlich der Kriterien Plausibilität und Ähnlichkeit besondere Priorität eingeräumt (Jonkisz et al., 2012). Für die sprachliche Konstruktion wurden die von Rost (2004) formulierten Hinweise berücksichtigt, wobei aufgrund der heterogenen Zielgruppe insbesondere auf eine fachlich korrekte, aber trotzdem verständliche Sprache und einen möglichst geringen Textumfang Wert gelegt wurde.

Nach der Konstruktion eines großen Itempools durch Projektmitarbeiterinnen und mitarbeiter mit unterschiedlicher fachlicher Expertise sowie intensiven Diskussionen zu jedem Item im interdisziplinären Projektteam, wurden die Items in ersten Feldtests erprobt. Hierzu gehörten informelle Prä-Tests, bei denen die Items Personen aus der Zielgruppe zur Bearbeitung vorgelegt wurden. Darüber hinaus wurde mit allen Items ein *Cognitive Lab* durchgeführt. Anhand der Technik des lauten Denkens werden hierbei kognitive Prozesse und Strategien identifiziert, die zur Bearbeitung und Lösung der Aufgaben erforderlich sind (Terzer, Patzke & Upmeier zu Belzen, 2012).

Insgesamt wurden so 117 Items konstruiert. Davon entfielen 53 auf das Fachwissen Mathematik, 42 auf das mathematikdidaktische Wissen und 22 Items auf das pädagogische Wissen. Da der Itempool damit fast doppelt so viele Items umfasste als mit 62 Items für den Test letztlich intendiert (24 für Fachwissen Mathematik, 22 mathematikdidaktisches Wissen und 16 pädagogisches Wissen), blieb für die weiteren Phasen der Testzusammenstellung Spielraum, die Items mit der größten Inhaltsvalidität auszuwählen.

4.3 Stichprobe für die Inhaltsvalidierung

Expertinnen und Experten aus Wissenschaft *und* Praxis wurden für die Einschätzung der Items herangezogen, um zugleich den aktuellen Forschungsstand und die Anforderungen des Berufsalltags zu berücksichtigen. Als Praktiker wurden Erzieherinnen mit Leitungsfunktion und/oder langjähriger Berufserfahrung, Aus- und Fortbildende im Bereich der Frühpädagogik sowie Fachberaterinnen herangezogen. Die Experten aus der Wissenschaft sind in Forschung und Lehre der Frühpädagogik bzw. der elementaren mathematischen Bildung als Hochschullehrende oder wissenschaftlich Mitarbeitende tätig, wobei bei ihrer Auswahl darauf geachtet wurde, dass unterschiedliche paradigmatische Zugänge vertreten waren.

Die Expertinnen und Experten standen in keinem beruflichen oder persönlichen Verhältnis zum Forschungsprojekt und erfüllten damit die Forderung nach *externen* Experten. Um deren Expertise quantifizieren zu können, wurde die Berufserfahrung im aktuellen Beruf in Jahren erfragt. Darüber hinaus wurden die Praktikerinnen, die mathematikbezogene Items beurteilen sollten, nach zusätzlichen Qualifikationen im Bereich Mathematik befragt und die Wissenschaftler nach der Anzahl ihrer Publikationen in den letzten fünf Jahren im betreffenden Bereich (siehe Tab. 1). Die Wissenschaftler wiesen im Mittel eine geringere Berufserfahrung auf als die Praktiker, was auf das noch junge Forschungsfeld der Frühpädagogik zurückzuführen ist. Zudem stellte die Dauer der Berufserfahrung bei den Praktikerinnen ein Auswahlkriterium dar.

	Wissenschaftler			**Praktiker**		**Gesamt**
	N	*Berufserfahrung*	*Publikationen*	*N*	*Berufserfahrung*	*N*
Fachwissen Mathematik	7	M = 7.7 SD = 5.9	M = 18.4 SD = 5.7	–	–	7
Mathematikdidaktisches Wissen	6	M = 6.83 SD = 2.79	M = 19.00 SD = 5.44	2	M = 10.50 SD = 7.78	8
Pädagogisches Wissen	2	M = 5.00 SD = 2.83	M = 18.50 SD = 0.71	7	M = 24.71 SD = 6.37	9

Anmerkungen. N = Anzahl der Expertinnen und Experten, M = Mittelwert, SD = Standardabweichung.

Tab. 1: Nachweis einschlägiger Expertise für die Einschätzung der Inhaltsvalidität

4.4 Durchführung der Inhaltsvalidierung

Die Beurteilung aller 117 Items durch jede Expertin bzw. jeden Experten hätte eine hohe zeitliche Beanspruchung bedeutet. Da auch nicht alle Befragten über ausgewiesene Expertise in jedem der drei Wissensbereiche verfügten, haben die Wissenschaftler jeweils Items aus zwei Wissensdomänen, die Experten aus der Praxis jeweils Items aus einer Domäne bearbeitet. Den Expertinnen und Experten wurden die Items zugesandt, um folgende Fragen auf einer vierstufigen Skala (1 = gar nicht, 2 = eher nein, 3 = eher ja, 4 = voll und ganz) zu beantworten (vgl. Hartig, Frey & Jude, 2012): „Wird der Inhalt durch das Item optimal repräsentiert?“ und „Stellt dieses Item eine gute Repräsentation aller (theoretisch) möglichen Items dar?“. Außerdem bestand die Möglichkeit, in offener Form Anmerkungen und Kommentare zu jedem Item bzw. generelle Anmerkungen am Ende zu notieren. Das Vorgehen war damit an Prozeduren angelehnt, wie sie beispielsweise unter den Begriffen *performance* bzw. *criterion centrality* oder *content authenticity* bekannt sind (z. B. Rothman, Slattery, Vranek & Resnick, 2002; Achieve, 2003; Alderson et al., 2006).

Die Items wurden ausgedruckt und zusammen mit wesentlichen Informationen über das Projekt und seine Ziele zugesandt. Zusätzliche Hintergrundinformationen zu den Inhalten der einzelnen Items wurden nicht mitgeteilt, da zum einen davon ausgegangen wurde, dass Experten per definitionem genügend Hintergrundwissen zu den jeweiligen Inhalten haben, und da es sich zum anderen um grundlegende Inhalte der Frühpädagogik handelte, zu denen konkrete Definitionen bzw. Vorstellungen existieren. Die Beantwortung der Fragen mit Raum für Anmerkungen erfolgte direkt im Anschluss an das Lesen jedes einzelnen Items.

4.5 Auswertung und Ergebnisse der Inhaltsvalidierung

Die Auswertung erfolgte quantitativ und qualitativ, wobei die Schritte eng vernetzt waren. Im quantitativen Schritt wurden die Items anhand des Mittelwertes der Experteneinschätzungen beurteilt, während im qualitativen Schritt eine Analyse der Kommentare erfolgte. Herangezogen wurde der arithmetische Mittelwert, der vor dem Hintergrund der geringen Stichprobengröße und durchgehend rechtssteiler Häufigkeitsverteilungen der Beurteilungswerte ein konservatives (i. e. strengeres) Kriterium der durchschnittlichen Einschätzung der Experten darstellt, da tendenziell eher niedrigere Mittelwerte resultierten. Andere Kennwerte wie Median oder Modus hätten bei solchen nicht-normalverteilten Daten zu einer liberaleren Einschätzung (also höheren, i. e. positiveren Werten) geführt, was bewusst nicht intendiert war. Darüber hinaus berücksichtigt der Mittelwert durch eine Gleichgewichtung aller Werte alle wissenschaftlichen und praktischen Experten gleichmäßig, was mit der Vorstellung von Expertise im Projekt übereinstimmt. Da in der Literatur kein Maß für die Auswertung von Expertenbefragungen gefunden werden konnte, wurde ein dreistufiges Schema entwickelt, mithilfe dessen die Items zugeordnet wurden (vgl. Abb. 2).

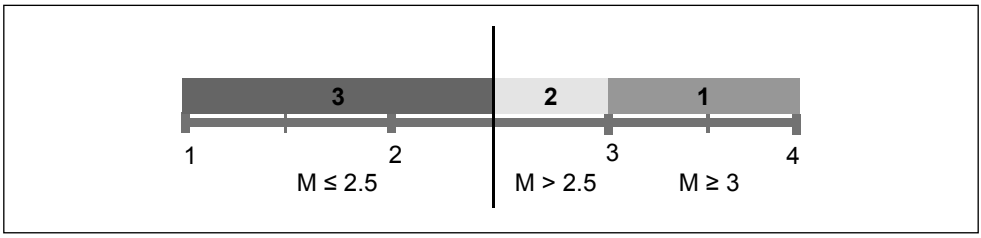


Abb. 2: Auswertungsschema zur Inhaltsvalidität in KomMa

Items, deren Mittelwert bei den Experteneinschätzungen (von 1 = gar nicht bis 4 = voll und ganz) auf einer der beiden Fragen kleiner oder gleich 2.5 war, wurden eliminiert, weil über die Hälfte der Experten das Item als eher oder gar nicht repräsentativ eingestuft hatte. Items, die auf beiden Fragen mindestens einen Mittelwert von 2.5 und auf einer Frage einen Mittelwert kleiner 3 aufwiesen, wurden nach einer umfassenden Revision beibehalten, für die die – anhand der einschlägigen Fachliteratur überprüften – Anmerkungen der Experten herangezogen wurden. Falls keine aussagekräftigen Anmerkungen vorhanden waren bzw. sich kein Konsens finden ließ, wurden auch diese Items eliminiert. Items, die auf beiden Fragen einen Mittelwert gleich oder größer 3 aufwiesen, wurden beibehalten, ggf. nach kleineren Verbesserungen aufgrund von plausiblen Expertenanmerkungen.

Insgesamt wiesen die Einschätzungen der Experten aus wissenschaftlicher und praktischer Sicht auf eine gute Inhaltsvalidität im Sinne einer hohen Repräsentativität der meisten Items für die zu erfassenden Konstrukte hin (siehe Tab. 3), sodass nur ein geringer Anteil eliminiert werden musste. Die Anmerkungen, die von den Experten zu revisionsbedürftigen Aspekten gemacht wurden, bezogen sich primär auf inhaltliche, aber auch auf sprachliche Aspekte der Items. Ein geringerer Teil der Anmerkungen bezog sich auf Praxiserfahrungen mit einem bestimmten Inhalt. Schwierigkeiten zeigten sich insbesondere bei der Einschätzung der offenen Items. Wenngleich den Items selbst oftmals sehr hohe Inhaltsvalidität bescheinigt wurde, musste diese durch die Kodieranwei-

N = 117	**n**	**Items eliminiert**	**Items revidiert**	**Items angenommen**
Fachwissen Mathematik	53	5	19	29
Mathematikdidaktisches Wissen	42	3	11	28
Pädagogisches Wissen	22	1	7	14
Gesamt	117	9	37	71

Anmerkungen. n = Anzahl der Items.

Tab. 2: Ergebnisse der Expertenbefragung

sungen ggf. wieder eingeschränkt werden, da die Zuordnung von richtigen und falschen Antworten nicht immer eindeutig nachvollziehbar war.

In jeder Dimension konnte die Mehrheit der Items ohne Beanstandungen angenommen werden. In den bei Lehrerkompetenzmessungen üblicherweise als schwierig geltenden Bereichen der Testkonstruktion, mathematikdidaktisches Wissen und pädagogisches Wissen, galt dies für rund zwei Drittel der Items, während die Items aus dem bei Lehrerkompetenzmessungen üblicherweise einfacher zu konstruierenden Bereich des Fachwissens Mathematik in KomMa stärker kritisiert wurden. Dieses Ergebnis lässt sich vermutlich darauf zurückführen, dass die Notwendigkeit mathematischen Fachwissens für die Frühpädagogik kontroverser beurteilt wird als für Lehrkräfte.

5. Zusammenfassung und Diskussion der gewonnenen Erkenntnisse sowie Ableitung von Empfehlungen

Ziel des vorliegenden Beitrags war, die vorhandenen Standards für die Validierung speziell von Kompetenztests zusammenzufassen. Zum anderen zielte der Beitrag darauf, konkrete Empfehlungen zur Sicherung von Inhaltsvalidität für die Testentwicklung bzw. Veröffentlichungspraxis vorzustellen, weil entsprechende Arbeiten bisher weitgehend fehlten, und diese beispielhaft empirisch zu illustrieren. In Tabelle 3 werden zunächst die verschiedenen Validierungsstrategien zu allen vorgestellten Validitätsaspekten systematisch zusammengefasst, bevor anschließend die gewonnenen Erkenntnisse zur Inhaltsvalidierung diskutiert werden.

Das oben im Detail vorgestellte Vorgehen der systematischen Expertenbefragung zur Inhaltsvalidierung in KomMa erwies sich als ökonomisch und zugleich erkennt-

Validitätsaspekt	**empfohlene Validierungsstrategien**
Inhaltsvalidität	systematische bzw. standardisierte Expertenbefragungen mit Nachweis der Expertise aller Beurteilerinnen und Beurteiler
Kriteriumsvalidität	Vorhersage von geeigneten Außenkriterien durch diagnostische Verfahren, wobei beide Verfahren zur selben Zeit eingesetzt werden (konkurrente Validität)
	Vorhersage von geeigneten Außenkriterien durch diagnostische Verfahren, wobei das Kriterium zu einem späteren Zeitpunkt erhoben wurde (prognostische Validität)
Konstruktvalidität	Prüfung theoretisch angenommener Strukturen mithilfe von konfirmatorischen Faktorenanalysen oder Modellen der Item-Response-Theorie (faktorielle Validität)
	Vorhersage von Itemschwierigkeiten anhand schwierigkeitsbestimmender Merkmale mithilfe von Regressionsanalysen oder erklärenden Modellen der Item-Response-Theorie (Prüfung der Konstruktrepräsentation)
	Prüfung von theoretisch angenommenen Zusammenhängen zu anderen Variablen mithilfe von Strukturgleichungsmodellen bzw. Modellen der Item-Response-Theorie (Prüfung eines nomologischen Netzes)

Tab. 3: Zusammenfassende Darstellung der Validitätsaspekte und ihrer jeweiligen Validierungsstrategien

nisgewinnend. Es empfiehlt sich besonders, wenn eine systematische Erfassung der Inhaltsvalidität auf Itemebene im Mittelpunkt des Interesses steht, wobei es sich auch auf die Erfassung von Inhaltsvalidität auf Testebene übertragen lässt. Die Möglichkeit, Anmerkungen der Expertinnen und Experten zu den einzelnen Items zu erfassen, erlaubt eine qualitative Einordnung der konstruierten Items in den Sachgegenstand und liefert gezielt Hinweise für die Überarbeitung und Neukonstruktion.

Bei der Durchführung und Auswertung der Expertenbefragung haben sich fünf Schwierigkeiten herausgestellt: (a) Beurteilungsfehler durch die gleiche (falsche) Vorstellung des Konstrukts von Testkonstrukteuren und Experten; (b) einseitige Fehlvorstellung des Konstrukts aufseiten der Testkonstrukteure oder Experten; (c) Transparenz von Projekthintergrund und -vorgehen; (d) Präzision der Kodieranweisungen und (e) Einbindung von Experten in weitere Phasen der Testkonstruktion.

Die Problematik von Beurteilungsfehlern ist aus der Kognitionspsychologie bekannt und wurde auch für die Urteile von Experten gezeigt. Deswegen verweist Kane (2013) darauf, dass Beurteilungsfehler bei der Erfassung von Inhaltsvalidität bedacht werden müssen. Als gravierendster Beurteilungsfehler muss die gemeinsame (Fehl)Vorstellung des Konstrukts durch Testkonstrukteure und Experten gesehen werden (a). Als Fehlvorstellung wurde in KomMa eine subjektive Überzeugung angesehen, die nicht mit dem aktuellen Stand allgemein akzeptierter Konzeptionen der betreffenden Domäne übereinstimmt (im Projekt KomMa die Frühpädagogik), wobei dieser allgemein akzeptierte Stand über einschlägige Fachliteratur definiert wurde. Diese Art des Beurteilungsfehlers führt dazu, dass zunächst nicht-inhaltsvalide Items konstruiert werden, denen dann Inhaltsvalidität bescheinigt wird. Somit würde der Test unter Umständen Aufgaben enthalten, die inhaltliche oder logische Fehler beinhalten.

Ein Weg, mit dieser Problematik umzugehen, ist, eine heterogene (und trotzdem qualifizierte) Gruppe von Expertinnen und Experten zu befragen. In KomMa wurden daher sowohl Praktiker als auch Wissenschaftler herangezogen, um eine einseitige Beurteilung zu vermeiden. Beide Gruppen betrachten das Konstrukt per definitionem aus unterschiedlichen Perspektiven: einer theoretisch-forschungsorientierten oder einer anwendungsorientierten Perspektive. Im vorliegenden Anwendungsfall wurde zudem ausdrücklich darauf hingewiesen, dass der Test bei Erzieherinnen während der Ausbildung eingesetzt werden soll und er nicht zum Zwecke der Individualdiagnostik konstruiert wurde. So sollte beispielsweise verhindert werden, dass die Praktiker Anforderungen erwarteten bzw. formulierten, die eher für berufserfahrene Erzieherinnen gelten. Zudem sei empfohlen, die Stichprobe des Expertenpanels hinreichend groß zu wählen, damit die Wahrscheinlichkeit der Identifikation von Fehlvorstellungen steigt.

Besteht eine einseitige Fehlvorstellung aufseiten der Testkonstrukteure (b), wird das Item revidiert oder eliminiert. Die Erfahrung der hier vorgestellten Expertenbefragung hat gezeigt, dass sich solche Fehlvorstellungen bereits in der quantitativen Beurteilung der Items zeigen. Die qualitativen Anmerkungen können dann herangezogen werden, das Item zu überarbeiten. Eine einseitige Fehlvorstellung aufseiten der Expertinnen und Experten zeigt sich z. B. in einer falschen Lösung des Items oder in Anmerkungen, die nicht zum Konsens der übrigen Experten passen bzw. sich anhand von Fachliteratur wi-

derlegen lassen. Da Kompetenztests in der Regel breite Konstrukte erfassen, muss allen Experten zugestanden werden, nicht überall einschlägig urteilen zu können. Sollte sich jedoch zeigen, dass ein Experte über mehrere Items hinweg systematisch falsch urteilt, sollte dieser aus der Auswertung ausgeschlossen werden. In der hier präsentierten Expertenbefragung war dies nicht erforderlich.

Diese Überlegungen machen aber deutlich, wie wichtig ein Nachweis einschlägiger Expertise ist (Jonson & Plake, 1998; Hornke & Winterfeld, 2004). Dies wird in den meisten uns bekannten Studien vernachlässigt. Als Kriterien für den Nachweis in KomMa wurde zum einen die Berufserfahrung gewählt, da langjährige Erfahrung in einer Domäne als Indikator für Expertise angesehen werden kann (Glaser, 1990). Für Wissenschaftler wurde zum anderen die Anzahl der Publikationen im betreffenden Bereich und bei Praktikern spezifische Angebote, die sie in ihrer Praxis im betreffenden Bereich durchgeführt haben, bzw. wahrgenommene Weiterbildungen erfragt. Ein weiterer Indikator könnte auch die Spezifität der Inhalte eines Berufs sein (Tesluk & Jacobs, 1998). Gezeigt hat sich, dass es einfacher ist, Experten für spezifische Teile des Tests einzusetzen, da die Gefahr von Fehlvorstellungen sinkt, während Experten bei einer Beurteilung des gesamten Konstrukts aufgrund von dessen Breite stärker der Gefahr von Fehlvorstellungen unterliegen. In KomMa erfolgte daher eine Eingrenzung auf eine bzw. zwei Wissensfacetten.

Ferner stellt sich die Frage, welche Kontextinformationen vor der Beurteilung der Items gegeben werden müssen (c). Je mehr Informationen den Expertinnen und Experten zur Verfügung stehen, desto valider gelingt ihnen die Beurteilung der Items (Pant, Rupp, Tiffin-Richards & Köller, 2009). In KomMa wurden daher zunächst wesentliche Hinweise zur Itemkonstruktion und zur Ableitung der Iteminhalte gegeben, da ein Großteil des Expertenpanels keine Erfahrung mit der empirischen Erfassung von Kompetenzen oder der Itemkonstruktion hatte und auf diesem Weg mit dem Format vertraut gemacht wurde, sodass später nicht die Art der Items, sondern deren Inhalt beurteilt werden konnte. Diesem Anschreiben wurden auch Beispielitems mit Erklärungen angefügt.

Eine besondere Problematik zeigte sich bei offenen Items (d), deren Antworten für die weitere Auswertung als richtig oder falsch beurteilt werden müssen. Hierzu wurden Kodieranweisungen entwickelt, die den Experten zusammen mit den Items vorgelegt wurden. Ihre Uneindeutigkeit wurde vielfach kritisch beurteilt, sodass auf dieses Problem vor zukünftigen Expertenbefragungen stärker geachtet werden sollte.

Über die Einbindung von Experten in die Beurteilung von Inhaltsvalidität auf Itemebene hinaus erscheint es sinnvoll, diese an anderen Stellen des Entwicklungsprozesses einzusetzen (e). Dies kann sich sowohl auf frühere Phasen, wie beispielsweise die Modellbildung, als auch auf spätere Phasen, wie beispielswiese die abschließende Testzusammenstellung, beziehen. Eine frühe Einbindung bietet sich vor allem deswegen an, da Kompetenztests oft breite Berufsfelder abdecken, zu denen bislang wenig empirische Erkenntnisse vorliegen (Blömeke & Zlatkin-Troitschanskaia, 2013). Die Einbindung von Experten in die Modellbildung kann bereits die inhaltliche Validität des Modells erhöhen, was sich positiv auf die Inhaltsvalidität des Tests auswirken sollte. Sollten die

ersten empirischen Überprüfungen der Items, z. B. im Rahmen einer Pilotierung, die Konstruktion von weiteren Items erfordern, können auch an dieser Stelle Experten eingesetzt werden.

Bei der Einbindung von Expertinnen und Experten in weitere Phasen der Testkonstruktion muss allerdings die Unabhängigkeit dieser gewahrt werden. Dementsprechend sollte auf eine klare Trennung von externen Empfehlungen und aktiver Mitarbeit geachtet werden. Außerdem ist eine Rotation der Experten in den einzelnen Phasen denkbar, um eine Zirkularität der Urteile zu vermeiden, indem in verschiedenen Phasen über die eigene Vorarbeit geurteilt wird. Hier bietet sich daher der Einsatz verschiedener Experten an. Voraussetzung hierfür ist allerdings wiederum, dass eine ausreichend große Anzahl an Experten verfügbar ist, was ggf. insbesondere bei einem breiten Spektrum an heranzuziehender Expertise zum Problem werden könnte.

Ein Test, der im Sinne von Kane (2013) inhaltsvalide Testwertinterpretationen zulässt und reliabel ist, liefert die Grundlagen für Konstrukt- und Kriteriumsvalidierungen, die für Kompetenztests unerlässlich sind. Der hier vorgestellte Ansatz der Inhaltsvalidierung bietet im Vergleich zu anderen Verfahren die Möglichkeit, die Validierung auf ökonomischem Weg in laufende Projekte zu integrieren. Den Nutzen dieses Ansatzes gilt es nun zu diskutieren und in weiteren Projekten auch im Vergleich zu Ansätzen wie der Generalizability-Theorie oder Inter-Rater-Reliabilität auf seine Tauglichkeit zu prüfen. Zudem wäre es wünschenswert, wenn ein breiterer Diskurs zu Standards der Inhaltsvalidierung begonnen würde, da dieser aus Sicht der Erziehungswissenschaft und der Fachdidaktiken mit Blick auf die Überzeugungskraft eines Kompetenztests besondere Bedeutung zukommt. Ein solcher Diskurs sollte Teilnehmerinnen und Teilnehmer unterschiedlicher Expertisebereiche einschließen, um zu ähnlich stark formalisierten Standards zu kommen, wie sie für die übrigen Validierungsstrategien schon lange gelten. Unser Beitrag sollte hierfür einen ersten Anstoß liefern.

Literatur

Achieve, Inc. (2003). *Review of Michigan's Grade-Level Content Expectations*. http://www.achieve.org/files/MI-FullReport10-05-04.pdf [09. 04. 2014].

AERA, APA & NCME – American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing*. Washington, D. C.: American Psychological Association.

AERA, APA & NCME – American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, D. C.: American Psychological Association.

Alderson, J. C., Figueras, N., Nold, G., North, B., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly, 3*, 3–30.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Hrsg.), *Test validity* (S. 9–13). Hillsdale: Lawrence Erlbaum.

APA American Psychological Association (1954). *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Washington, D. C.: American Psychological Association.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Schümer, G., Stanat, P., Tillmann, K.-J., & Weiß, M. (2003). *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Zusammenfassung zentraler Befunde*. Berlin: Max-Planck-Institut für Bildungsforschung.

Blömeke, S. (2013). *Validierung als Aufgabe im Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor"* (KoKoHS Working Papers, 2). Berlin/Mainz: Humboldt-Universität/Johannes Gutenberg-Universität.

Blömeke, S., Bremerich-Vos, A., Haudeck, H., Kaiser, G., Lehmann, R., Nold, G., Schwippert, K., & Willenberg, H. (Hrsg.) (2011). *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen: Erste Ergebnisse aus TEDS-LT*. Münster: Waxmann.

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

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

Blömeke, S., & Zlatkin-Troitschanskaia, O. (2013). *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs* (KoKoHs Working Papers, 1). Berlin/Mainz: Humboldt-Universität/Johannes Gutenberg-Universität.

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

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Aufl.). Heidelberg: Springer.

Brunner, M., Kunter, M., Krauss, S., Klusmann, U., Baumert, J., Blum, W., et al. (2006). Die professionelle Kompetenz von Mathematiklehrkräften: Konzeptualisierung, Erfassung und Bedeutung für den Unterricht. Eine Zwischenbilanz des COACTIV-Projekts. In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (S. 54–82). Münster: Waxmann.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Dunekacke, S., Jenßen, L., Baack, W., Tengler, M., Wedekind, H., Grassmann, M., & Blömeke, S. (2013). Was zeichnet eine kompetente pädagogische Fachkraft im Bereich Mathematik aus? Modellierung professioneller Kompetenz für den Elementarbereich. *Beiträge zum Mathematikunterricht, 2013*, 280–283.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Frey, A. (2013). *Validität.* Eröffnungsvortrag im Rahmen des KoKoHs-Rundgesprächs zu Validität und Validierung am 14. März 2013 an der Humboldt-Universität zu Berlin.

Fried, L., & Roux, S. (2009). Zur Pädagogik der frühen Kindheit im 21. Jahrhundert – Desiderata. In L. Fried & S. Roux (Hrsg.), *Pädagogik der frühen Kindheit. Handbuch und Nachschlagewerk* (2. Aufl., S. 378–382). Berlin: Cornelsen.

Gärtner, H., & Pant, H. A. (2011). Validity of processes and results of school inspection. *Studies in Educational Evaluation, 37*(2-3), 85–93.

Geisinger, K. F. (1992). The metamorphosis in test validation. *Educational Psychologist, 27*, 197–222.

Glaser, R. (1990). Expertise. In M. W. Eysenk, A. N. Ellis, E. Hunt & P. Johnson-Laird (Hrsg.), *The Blackwell dictionary of cognitive psychology* (S. 139–142). Oxford: Blackwell Reference.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1–10.

Hartig, J. (2013). Workshop „Konstruktvalidität" im Rahmen des KoKoHs-Rundgesprächs zu Validität und Validierung am 14./15. März 2013 an der Humboldt-Universität zu Berlin.

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau, 63*, 43–49.

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 143–171). Heidelberg: Springer.

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72*, 665–686.

Hornke, L. F., & Winterfeld, U. (2004). *Eignungsbeurteilungen auf dem Prüfstand: DIN 33430 zur Qualitätssicherung*. Heidelberg: Spektrum.

Jonkisz, E., Moosbrugger, H., & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 27–74). Heidelberg: Springer.

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*, 736–753.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kersting, N. (2008). Using Video Clips of Mathematics Classroom Instruction as Item Prompts to Measure Teachers' Knowledge of Teaching Mathematics. *Educational and Psychological Measurement, 68*, 845–861.

Klauer, K. J. (1984). Kontentvalidität. *Diagnostica, 30*, 1–23.

Kubinger, K. D. (2006). *Psychologische Diagnostik*. Göttingen: Hogrefe.

Kunina-Habenicht, O., Lohse-Bossenz, H., Kunter, M., Dicke, T., Förster, D., Gößling, J., Schulze-Stocker, F., Schmeck, A., Baumert, J., Leutner, D., & Terhart, E. (2012). Welche bildungswissenschaftlichen Inhalte sind wichtig in der Lehrerbildung? Ergebnisse einer Delphi-Studie. *Zeitschrift für Erziehungswissenschaft, 15*(4), 649–682.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Lohse-Bossenz, H., Kunina-Habenicht, O., & Kunter, M. (2013). The role of educational psychology in teacher education: expert opinions on what teachers should know about learning, development, and assessment. *European Journal of Psychology of Education, 28*, 1543–1565.

Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational Measurement* (3. Aufl., S. 13–104). New York: American Council on Education/Macmillan.

Messick, S. (1995). Validity of Psychological Assessment. Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Scoring Meaning. *American Psychologist, 50*(9), 741–749.

National Mathematics Advisory Panel (2008). *The Final Report of the National Mathematics Advisory Panel*. Washington, D. C.: Department of Education.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35*, 95–101.

Pauli, C., & Reusser, K. (2006). Von international vergleichenden Video Surveys zur videobasierten Unterrichtsforschung und -entwicklung. *Zeitschrift für Pädagogik, 52*(6), 774–797.

Popham, W. J. (1993). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*, 285–301.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? *Psychologische Rundschau, 57*(2), 69–86.

Rossiter, J. R. (2008). Content Validity of Measures of Abstract Constructs in Management and Organizational Research. *British Journal of Management, 19*, 380–388.

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

Roth, X. (2013). Quereinstiege – Eine ressourcenorientierte Betrachtung. *Frühe Bildung, 2*(2), 92–97.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and Alignment of Standards and Testing*. http://cse.ucla.edu/products/reports/TR566.pdf [03. 03. 2014].

Schaper, N. (2013). Workshop „Externe Validität" im Rahmen des KoKoHs-Rundgesprächs zu Validität und Validierung am 14./15. März 2013 an der Humboldt-Universität zu Berlin.

Scriven, M. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31*(1), 105–117.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*(2), 4–14.

Speck-Hamdan, A. (2011). *Grundschulpädagogisches Wissen – Impulse für die Elementarpädagogik? Eine Expertise der Weiterbildungsinitiative Frühpädagogische Fachkräfte (WiFF)*. München: Deutsches Jugendinstitut.

Terzer, E., Patzke, C., & Upmeier zu Belzen, A. (2012). Validierung von Multiple-Choice Items zur Modellkompetenz durch lautes Denken. In U. Harms & F. X. Bogner (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik* (S. 45–62). Innsbruck: Studienverlag.

Tesluk, P. E., & Jacobs, R. R. (1998). Toward an integrated model of work experience. *Personnel Psychology, 51*, 321–355.

Watermann, R., & Klieme, E. (2002). Reporting Results of Large-Scale Assessment in Psychologically and Educationally Meaningful Terms. Construct Validation and Proficiency Scaling in TIMSS. *European Journal of Psychological Assessment, 18*(3), 190–203.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.

Wottawa, H., & Hossiep, R. (1997). *Anwendungsfelder psychologischer Diagnostik*. Göttingen: Hogrefe.

Yalow, E. S., & Popham, W. J. (1983). Content Validity at the Crossroads. *Educational Researcher, 12*, 10–14.

Zwick, R., & Himelfarb, I. (2011). The Effect of High School Socioeconomic Status on the Predictive Validity of SAT Scores and High School Grade-Point Average. *Journal of Educational Measurement, 48*, 101–121.

Abstract: The article discusses requirements of validation approaches with respect to competence assessments with a special focus on how to provide evidence of content validity. As part of a validation framework that covers the range of validation approaches necessary, a procedure for collecting evidence on content validity in the context of a competence test is introduced by using the research project *KomMa* as an example. External research experts and practitioners helped to conduct an efficient rating in written form that covered the content quality of the items developed. Step-by-step, the validation procedure which was developed in the context of KomMa is outlined and put up for discussion. The article concludes with offering practical tips for the implementation of such a content validation procedure including recommendations for additional validation strategies organized according to the different facets of validity.

Keywords: Validity, Test, Content Validation, Expert Rating, Competencies

Anschrift des Autors/der Autorinnen

Dipl.-Psych. Lars Jenßen, Humboldt-Universität zu Berlin,
Institut für Erziehungswissenschaften, Abteilung Systematische Didaktik
und Unterrichtsforschung, Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: lars.jenssen@hu-berlin.de

M.A. Simone Dunekacke, Humboldt-Universität zu Berlin,
Institut für Erziehungswissenschaften, Abteilung Systematische Didaktik
und Unterrichtsforschung, und Carl von Ossietzky Universität Oldenburg,
Institut für Pädagogik, Uhlhornsweg, 26111 Oldenburg, Deutschland
E-Mail: simone.dunekacke@uni-oldenburg.de

Prof. Dr. Sigrid Blömeke, Centre for Educational Measurement at the University of Oslo
(CEMO), Leibniz-Institut für die Pädagogik der Mathematik und Naturwissenschaften Kiel,
Humboldt-Universität zu Berlin, Postboks 1072/Blindern, 0316 Oslo, Norwegen
E-Mail: sigribl@cemo.uio.no

Svenja Hammer/Sonja A. Carlson/Timo Ehmke/Barbara Koch-Priewe/
Anne Köker/Udo Ohm/Sonja Rosenbrock/Nina Schulze

Kompetenz von Lehramtsstudierenden in Deutsch als Zweitsprache

Validierung des GSL-Testinstruments

Zusammenfassung: Fachlehrkräfte benötigen in der Lehrerausbildung Gelegenheit zum Kompetenzerwerb in Deutsch als Zweitsprache (DaZ), um die spezifischen Lernvoraussetzungen von Schülerinnen und Schülern nicht-deutscher Herkunftssprache berücksichtigen zu können. Auf der Basis eines Kompetenzmodells mit den Dimensionen *Fachregister*, *Mehrsprachigkeit* und *Didaktik* wurde im BMBF-Projekt *DaZKom* derzeit ein Testinstrument entwickelt und mit IRT-Methoden ausgewertet. Die Ergebnisse der Validierungsstudie (N = 252) zeigen, dass Zusammenhänge zwischen DaZ-Kompetenz und linguistischem Wissen sowie pädagogischem Wissen bestehen, der Test aber konzeptuell ein eigenständiges Konstrukt misst. Darüber hinaus korrelieren eine höhere Semesterzahl, Deutsch als Studienfach und eine größere Anzahl an DaZ-Lerngelegenheiten mit einer höheren DaZ-Kompetenz.

Schlagworte: Deutsch als Zweitsprache, Lehrerbildung, Kompetenz, Validierung, Fachdidaktik Mathematik

1. Einleitung

Zahlreiche Schulleistungsstudien haben signifikante Unterschiede zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund in schulischen Leistungen aufgezeigt (OECD, 2013a; Blossfeld et al., 2007). Dass ein Zusammenhang zwischen Schulerfolg und ausreichender Kompetenz in der Unterrichtssprache Deutsch sowie anderen Faktoren wie dem sozioökonomischen Hintergrund besteht, kann daher als unbestritten gelten. Die Notwendigkeit systematischer Sprachförderung von leistungsschwachen Schülergruppen in jedem regulären Fachunterricht steht im bildungswissenschaftlichen Diskurs deshalb mittlerweile außer Frage (Leung, 2007; Gibbons, 2002; Schleppegrell, 2004, 2007; Prediger & Özdil, 2011; Ohm, 2009, 2010; Scarcella, 2003; Shanahan & Shanahan, 2008). Es gibt jedoch bislang kaum empirische Studien darüber, wie fach-

übergreifende Förderung im Vergleich zu fach- und DaZ-spezifischer fachintegrierter Förderung wirkt (Echevarria, Short & Vogt, 2008).

Fachlehrkräfte sind zudem für diese Aufgabe i. d. R. nicht angemessen ausgebildet und fühlen sich häufig für Maßnahmen zur sprachlichen Förderung auch nicht zuständig (Becker-Mrotzek, Hentschel, Hippmann & Linnemann, 2012; de Jong, 2013; Li & Zhang, 2004; Simmons, 2009; Zhang & Stephens, 2013). Alle Fachdidaktiken sollten daher in der universitären Lehrerausbildung auch auf die jeweiligen sprachlichen Interventionen vorbereiten, die das Handeln im Fachunterricht aller Fächer erfordert (Commins & Miramontes, 2006; de Jong & Harper, 2005). Trotz bislang fehlender empirischer Belege kann die Verbindung von fachdidaktischer Professionalität mit Kompetenzen im Bereich Deutsch als Zweitsprache als *generische* Kompetenz angesehen werden. Wenn – wie es zurzeit an vielen universitären Standorten geschieht – Inhalte des Fachs *Deutsch als Zweitsprache* in die reguläre Lehrerausbildung integriert werden sollen, ist es notwendig, sich über Standards bzw. Niveaus zu verständigen (Lucas & Grinberg, 2008) und dann auch zu prüfen, welche Lerngelegenheiten die Universitäten bieten können bzw. sollten, in denen die entsprechenden Kompetenzen erworben werden können. Im Rahmen des vom BMBF geförderten Projekts *Professionelle Kompetenzen angehender Lehrerinnen und Lehrer (Sek I) im Bereich ‚Deutsch als Zweitsprache' (DaZKom)* ist auf Basis eines theoretisch fundierten Kompetenzmodells das GSL-Testinstrument[1] entwickelt worden, mit dem DaZ-Kompetenz von Studierenden empirisch erfasst werden kann und sich Zusammenhänge mit universitären Lerngelegenheiten im Bereich DaZ untersuchen lassen.

Der vorliegende Beitrag konzentriert sich vor allem auf Fragen der Validierung und berücksichtigt dabei die Entstehung sowie die Überprüfung des oben genannten Kompetenzmodells. Die zentrale Fragestellung dieser Studie lautet: Testet der GSL-Test eine Kompetenz, die mit bereits eingeführten Testinstrumenten angrenzender Kompetenzdomänen nicht oder jedenfalls nicht hinreichend erfasst werden kann? Zudem sollen Zusammenhänge zwischen der DaZ-Kompetenz und Personenmerkmalen bzw. der Nutzung von DaZ-bezogenen Lerngelegenheiten als externe Validitätskriterien herangezogen werden.

Im Folgenden werden zunächst Generierung und Rahmenkonzeption des Kompetenzmodells umrissen (Abschnitt 2). Anschließend werden der argumentationsbasierte Validierungsansatz von Kane (2013) und die im Beitrag vorgenommene Fokussierung vorgestellt (Abschnitt 3). Es schließt sich eine differenzierte Darstellung der Forschungsfragen und der ihnen zugrunde liegenden Forschungshypothesen an (Abschnitt 4). Dabei wird auch dargestellt, was für Kompetenzdimensionen und -facetten die jeweiligen Testinstrumente angrenzender Kompetenzdomänen in der Lage sind zu erfassen und in welchem Verhältnis diese Dimensionen und Facetten zu denen des zu validierenden DaZ-Testinstruments stehen. Im Kapitel zum methodischen Vorgehen wird neben den notwendigen Angaben zum Erhebungs- und Auswertungsverfahren ausführlich die Operationalisierung der Fragestellung mithilfe der externen Testkonstruk-

1 GSL = German as a Second Language.

te beschrieben (Abschnitt 5). Es folgt eine zusammenfassende Darstellung der Ergebnisse für die unterschiedlichen Erhebungsbereiche (Abschnitt 6). Der Beitrag schließt mit einer Diskussion, in der Grenzen und Konsequenzen für die Weiterentwicklung des DaZ-Testinstruments, die Grenzen der Untersuchung sowie Forschungsdesiderata thematisiert werden (Abschnitt 7).

2. Konzeptualisierung von DaZ-Kompetenz

Das Projekt *DaZKom* orientiert sich an den theoretischen Rahmenkonzeptionen von Studien wie *Mathematics Teaching in the 21st Century* (MT21; Schmidt et al., 2007) oder *Teacher Education and Development Study: Learning to Teach Mathematics* (TEDS-M; Blömeke, Kaiser & Lehmann, 2010), die ebenfalls auf die Erfassung von Lehrerkompetenzen im Unterrichtsfach Mathematik abzielen. Es existierten jedoch zu Projektbeginn keine Vorarbeiten, die die generische Kompetenz Deutsch als Zweitsprache von Regelschullehrenden empirisch abgesichert modellieren und an die man hätte anschließen können. In Bezug auf sprachbezogene Kompetenzen von Erzieherinnen und Erziehern lieferte das Projekt *SprachKoPF* einige Vorarbeit (Hopp, Thoma & Tracy, 2010). Für die inhaltliche Rahmenkonzeption von DaZ-Kompetenz in der Schule bildete eine Dokumentenanalyse von 60 Curricula deutscher Universitäten u. a. der Fächer Deutsch als Fremd- und Zweitsprache die Basis. Das auf diese Weise generierte und wiederum durch eine erste Expertenbefragung bestätigte Rahmenkonzept (vgl. Köker et al., 2015) umfasst drei Kompetenzdimensionen mit inhaltlich ausformulierten Subdimensionen und Kompetenzfacetten. Abbildung 1 zeigt die Struktur von DaZ-Kompetenz in Dimensionen und Subdimensionen.

Die Unterteilung in Dimensionen mit den Ausdifferenzierungen in Subdimensionen und inhaltlichen Facetten ist analytisch zu verstehen. Im Unterrichtshandeln von Fachlehrkräften kommen immer alle Dimensionen zum Tragen, wenn auch in unterschiedlicher Gewichtung. Dieses Ineinandergreifen der Dimensionen soll im Folgenden konzeptionell umrissen werden.

Das Kompetenzmodell knüpft an den von Maas (2008) beschriebenen „in der Registerdifferenzierung verankerte[n] Typ von Mehrsprachigkeit“ (S. 53) an. Entgegen anderer Typen von Mehrsprachigkeit, die nicht selten ausbalancierte (z. B. erfolgreich bilingual aufwachsende Kinder von Eltern mit unterschiedlichen Erstsprachen) oder gar virtuose (z. B. professionelle Mehrsprachigkeit von Übersetzern) Formen der Beherrschung mehrerer Sprachen voraussetzen, ist dieser Typus funktional auf die Ausdifferenzierung des Sprachgebrauchs nach Registern bezogen.[2] Kinder wachsen im Laufe ihrer biografischen Entwicklung in unterschiedliche gesellschaftliche Domänen hin-

2 Der dem vorliegenden Beitrag zugrunde gelegte Registerbegriff stützt sich auf Halliday (1978, S. 195), der unter Sprachentwicklung eine Vergrößerung der Bandbreite der sozialen Funktionen von Sprache versteht. Dies schlägt sich in der Entwicklung neuer sprachlicher Register nieder. Den Registerbegriff definiert Halliday wie folgt: „A register is a set of mean-

	Dimension	Subdimension
DaZ-Kompetenz	Fachregister (Fokus auf Sprache)	Grammatische Strukturen und Wortschatz
		Semiotische Systeme
	Mehrsprachigkeit (Fokus auf Lernprozess)	Zweitspracherwerb
		Migration
	Didaktik (Fokus auf Lehrprozess)	Diagnose
		Förderung

Abb. 1: Kompetenzmodell für Deutsch als Zweitsprache

ein, die sich durch ihren Grad an Öffentlichkeit und Formalität unterscheiden (Familie, Freundeskreis, Kita, Geschäfte, Schule etc.). Diese Domänen müssen sie sich auch sprachlich erschließen. Hierbei kommen je nach Gesellschaft unter Umständen unterschiedliche Sprachen ins Spiel. Während beispielsweise in vielen mehrsprachigen afrikanischen Gesellschaften für das familiale, intime Register (Familie, Freundeskreis), das informelle öffentliche Register (öffentliches Leben auf der Straße, in Geschäften etc.) und das formelle Register (z. B. Schule, Ämter u. a. Institutionen) je unterschiedliche Sprachen gebraucht werden (vgl. Maas, 2008, S. 57), wird in der auf den ersten Blick monolingualen deutschen Gesellschaft für alle Register nur eine Sprache benötigt. Bei genauer Betrachtung werden jedoch für das intime Register und regional unterschiedlich ausgeprägt auch für das informelle öffentliche Register vielfach dialektale Formen des Deutschen gebraucht, während lediglich für das formelle Register einheitlich die deutsche Hochsprache verwendet wird. Insofern kann man für Deutschland zumindest eine *innere Mehrsprachigkeit* (Gebrauch mehrerer dialektaler Formen unter einer gemeinsamen Hochsprache) ansetzen.

---

ings that is appropriate to a particular function of language, together with the words and structures which express these meanings“ (ebd.).

Nimmt man die regionalen Minderheitensprachen (z. B. Dänisch, Friesisch, Sorbisch) und die vor allem in städtischen Räumen verstärkt auftretenden Migrantensprachen hinzu, lässt sich die Vorstellung einer monolingualen deutschen Gesellschaft schwerlich aufrechterhalten. Insbesondere die mit Blick auf DaZ-Kompetenz interessierenden migrantischen Sprecher verwenden nicht nur für das intime Register, sondern häufig auch für das informelle öffentliche Register ihre jeweilige Herkunftssprache. In der Regel erschließt sich aber insbesondere für die nachwachsenden Generationen die informelle Öffentlichkeit vollständig nur über den zusätzlichen Gebrauch des Deutschen (einschließlich unterschiedlicher Formen der Sprachmischung und des Sprachwechselns). Bereits die vorschulischen Bildungsinstitutionen (Kita, Kindergarten), spätestens aber die Schule markieren in der (bildungs-)biografischen Entwicklung den Übergang von der informellen zur formellen Öffentlichkeit mit entsprechenden Erwartungen bzw. Anforderungen an den Registergebrauch. So ist der Gebrauch des formellen öffentlichen Registers in der Bildungsinstitution Schule fast ausnahmslos mit der deutschen Hochsprache verbunden.

Die Dimension *Mehrsprachigkeit* des vorliegenden Kompetenzmodells bezeichnet das Wissen der angehenden Lehrkräfte über die angedeuteten Zusammenhänge zwischen individueller und gesellschaftlicher Mehrsprachigkeit (innerer wie äußerer) sowie die Fähigkeit, dieses Wissen für die Unterstützung zweitsprachlicher Lernprozesse mit Fokus auf die Entwicklung des für das Unterrichtsfach Mathematik typischen Fachregisters (siehe unten) zu nutzen. Während die Subdimension *Zweitspracherwerb* dabei auf den individuellen Erwerbsprozess sowie die Bedingungen, unter denen dieser abläuft, und die Faktoren, die seinen Verlauf beeinflussen, verweist, bezieht sich die Subdimension *Migration* auf die sprachliche Vielfalt in der Schule und den Umgang mit der dadurch im Fachunterricht entstehenden Heterogenität.

Bei der Skizzierung der Dimension *Mehrsprachigkeit* wurde bereits auf den in (bildungs-)biografischer Perspektive engen Zusammenhang mit der Registerdifferenzierung im Rahmen der Erschließung gesellschaftlicher Domänen eingegangen. Dabei stand aus der Perspektive der Mehrsprachigkeit der Erwerb weiterer Sprachen bzw. Varietäten im Vordergrund. Die Dimension *Fachregister* thematisiert nun das formelle öffentliche Register, dessen Gebrauch im Fachunterricht der Schule von allen Schülerinnen und Schülern erwartet wird. Es wird in der Hochsprache Deutsch realisiert und verlangt Formen der Bedeutungskonstruktion, der sprachlichen Argumentation und der Kombination von sprachlichen Elementen (vgl. die Registerdefinition von Halliday, 1978, in Fußnote 2), die sich deutlich von denen unterscheiden, die Schülerinnen und Schülern aus der Alltagskommunikation im intimen oder informellen öffentlichen Register geläufig sind. Für dieses Register hat sich in der deutschsprachigen Fachdiskussion der Begriff *Bildungssprache* etabliert, der allerdings eher auf übergreifende, nicht allein im schulischen Kontext vorkommende und damit auch „nicht auf einzelne schulische Lernbereiche oder Fächer bezogene Merkmale“ abzielt (Gogolin, 2009, S. 270). Gogolin charakterisiert Bildungssprache, indem sie ihre zentralen Merkmale denen der Alltagssprache gegenüberstellt: „Zusammenfassend und global charakterisiert, weist also *Bildungssprache* tendenziell die Merkmale formeller, mo-

nologischer schriftförmiger Kommunikation auf, während Alltagssprachgebrauch eher dialogisch gestaltet ist und die Merkmale informeller mündlicher Kommunikation aufweist“ (ebd.).

Das vorliegende Kompetenzmodell bezieht sich auf die DaZ-Kompetenz von angehenden Lehrkräften des Unterrichtsfaches Mathematik. Es fokussiert auf die spezifischen bildungssprachlichen Anforderungen dieses Fachs und beschäftigt sich daher insbesondere mit dem mathematischen Fachregister. Die Dimension *Fachregister* bezeichnet im derzeitigen Kompetenzmodell das Wissen der angehenden Lehrkräfte über dieses Register und ihre Fähigkeit, dieses Wissen für den Mathematikunterricht lernförderlich einzusetzen. Dabei knüpft die Unterteilung in die Subdimensionen *Grammatische Strukturen und Wortschatz* sowie *Semiotische Systeme* an Schleppegrell (2007) an, die darauf hinweist, dass Mathematik sich einer Vielzahl semiotischer Systeme bedient, mit denen Bedeutung erzeugt wird. *Fachregister* bezeichnet demnach nicht nur Sprache im Mathematikunterricht im engeren Sinne von Wortschatz und grammatischen Strukturen, sondern auch die Unterscheidung der unterschiedlichen semiotischen Systeme (mathematische Symbolnotation, mündliche und schriftliche Sprache, grafische und bildliche Darstellungen). Für die Subdimension *Grammatische Strukturen und Wortschatz* ist von grundlegender Bedeutung, dass sich deren fachregistertypische Merkmale im Unterricht medial sowohl schriftlich als auch mündlich realisieren, dass sie sich konzeptionell aber an schriftsprachlichen Erwartungen orientieren (zur Differenzierung Medialität/Konzeptionalität vgl. Koch & Oesterreicher, 1994).

Die Dimension *Didaktik* beschreibt schließlich die Teilkompetenz einer Lehrperson, den Lehrprozess im Fachunterricht auf der Basis der mit den Dimensionen *Mehrsprachigkeit* und *Fachregister* bezeichneten Teilkompetenzen sowohl in der unmittelbaren Unterrichtsinteraktion (Mikro-Scaffolding) als auch mit Blick auf die längerfristige Unterrichtsplanung (Makro-Scaffolding) sprachlernförderlich zu gestalten. Dazu müssen Lehrkräfte in der Lage sein, die sprachlichen Leistungen ihrer Schülerinnen und Schüler zu analysieren *(Diagnose)* und zu unterstützen *(Förderung)*.

Die Niveaubeschreibung des DaZ-Kompetenzmodells orientiert sich am Modell des Fertigkeitserwerbs von Dreyfus & Dreyfus (1986). Diese beschreiben die Kompetenzentwicklung als dynamischen Prozess, der in fünf Stufen verläuft und nur dann vom Novizen zum Experten führt, wenn der Lernende die Möglichkeit hat, ausreichend Erfahrungen zu machen (S. 20). Auch das DaZ-Kompetenzmodell versteht Kompetenzentwicklung nicht als rein kognitiven Prozess, sondern bezieht Erfahrungslernen und Handlungsorientierung mit ein. Es wurden zunächst drei Niveaustufen – vom Novizen über den fortgeschrittenen Anfänger bis zum kompetent Handelnden – modelliert und operationalisiert. Diese Stufung kann im Folgenden aus Platzgründen lediglich andeutungsweise erläutert werden (ausführlich in Köker et al., 2015).

Der *Novize* lernt, wie man Fakten und relevante Muster erkennt, und er lernt Regeln, mit denen er aufgrund der Fakten und Muster seine Handlungen bestimmen kann. Welche Fakten und Muster für sein Handeln als relevant anzusehen sind, ist für den Novizen so klar und objektiv definiert, dass er sie ohne Bezug auf die Gesamtsituation, in der sie auftauchen, d. h. kontextfrei bzw. kontextunabhängig, beurteilt (Dreyfus & Drey-

fus, 1986, S. 21). So hat eine angehende Fachlehrkraft gelernt, dass die Aneignung von Fachinhalten durch das Vorkommen unbekannter Wörter in Fachtexten, Übungen und Aufgaben erschwert wird. Aus diesem erlernten Faktum ergibt sich für die Lehrkraft die Regel, dass sie Fachtexte, die sie im Unterricht einsetzen will, auf unbekannte Wörter hin untersucht. Die Lehrkraft handelt hier kontextfrei, weil sie jede unterrichtliche Situation sprachlich nach dem Muster ‚ein unbekanntes Wort ist ein schwieriges Wort und deshalb eine relevante Lernschwierigkeit' beurteilt, ohne sich auf andere Elemente der Gesamtsituation zu beziehen.

Der *fortgeschrittene Anfänger* hat umfangreiche Erfahrungen darüber gesammelt, wie man mit „wirklichen" Situationen fertig wird. Dazu zählt vor allem, dass er in konkreten Situationen praktische Erfahrungen beim Umgang mit bedeutungsvollen Elementen gemacht hat, die nicht in objektiv fassbaren, kontextfreien Begriffen definiert werden können. Er erkennt solche situationsbezogenen Elemente, indem er zwischen ihnen und Beispielen aus früheren Erfahrungen Ähnlichkeiten wahrnimmt (Dreyfus & Dreyfus, 1986, S. 23). Der fortgeschrittene Anfänger kann sein Verhalten sowohl durch Bezugnahme auf die neuen, situationsbezogenen Elemente der Stufe II als auch auf die kontextfreien Fakten und Regeln der Stufe I steuern. Die o. g. angehende Lehrkraft hat beispielsweise in der Praxisphase ihres Lehramtsstudiums die Erfahrung gemacht, dass das Vorkommen unbekannter Wörter (vgl. Stufe I) kein allgemeingültiger Indikator für sprachliche Schwierigkeiten bei der Aneignung von Fachinhalten ist. Sie weiß nun, dass vermeintlich leichte, aus der Alltagssprache bekannte trennbare komplexe Verben (sog. Partikelverben) wie *abziehen* die sprachliche Komplexität auf syntaktischer Ebene erhöhen können, weil beim finiten Gebrauch Verb und Partikel eine Satzklammer bilden und daher getrennt stehen („Man *zieht* den Betrag, der sich aus der Berechnung ergibt, von der Summe *ab*.").

Da es für ihn mit zunehmender Erfahrung immer schwieriger wird, zu entscheiden, was in einer Situation beachtet werden muss, hat der *kompetent Handelnde* gelernt, hierarchisch geordnete Entscheidungsprozeduren anzuwenden. Er wählt zunächst einen Plan, um eine Situation zu organisieren, und untersucht dann nur noch die kleine Menge an Faktoren, die im betreffenden Plan am wichtigsten sind (Dreyfus & Dreyfus, 1986, S. 23–24). Das Vorgehen wird somit ganz entscheidend vom Handelnden selbst beeinflusst. Auf dieser Stufe plant die o. g. Lehrkraft den Einsatz von Fachtexten auf der Basis einer systematischen Analyse der sprachlichen Anforderungen aus der Fachregisterperspektive. Sie analysiert Schülerproduktionen systematisch und kontinuierlich mit Blick auf die erreichten Entwicklungsstufen und setzt diese in Beziehung zur fachlich notwendigen Ausdifferenzierung des Fachregisters. Die Lehrkraft stellt sich beispielsweise die Frage, wann die durch trennbare Verben erzeugte Satzklammer aus fachlicher Sicht rezeptiv bzw. produktiv beherrscht werden muss, welche Verben jeweils bekannt sein müssen, welche nicht-trennbaren bzw. synonymen Verben möglicherweise bereits bekannt sind („subtrahieren", „abziehen") und bezieht diese Überlegungen in ihre Planung mit ein.

Das Strukturmodell zur DaZ-Kompetenz ist als generisch zu betrachten und sollte bei entsprechender Anpassung an die jeweiligen fachregistertypischen Anforderungen

auf unterschiedliche Domänen oder Unterrichtsfächer angewendet werden können. Als inhaltliche Domäne wurde bei der Testentwicklung im Rahmen des DaZKom-Projekts das Bezugsfach Mathematik gewählt.

3. Validierungsansatz

Validierung bezeichnet im Folgenden einen Prozess, in dem belegt werden soll, dass das vorliegende Testinstrument seinen angestrebten Zweck erfüllt (Sireci & Padilla, 2014). Hierzu wird der argumentationsbasierte Ansatz *(argument-based approach)* von Kane (2013) genutzt. Bei diesem besteht das Ziel darin, größtmögliche Klarheit und Nachvollziehbarkeit in der Auslegung von Testergebnissen und deren Interpretationen zu erlangen (Kecker, 2011). Dafür wurde von Kane eine Argumentationskette entwickelt, die es ermöglicht, eine Verbindung zwischen der beobachteten Testleistung und der Ergebnisinterpretation bezogen auf die Realsituation herzustellen. Kane, Crooks und Cohen (1999) untergliedern diese Argumentationskette in drei Schritte: *evaluation, generalization* und *extrapolation.* Im vorliegenden Artikel konzentrieren wir uns auf den Schritt der *extrapolation* (Schlussfolgerung; Kane, 2013). Bei der Extrapolation wird das Testergebnis einer Person als Indikator für ihre zukünftige Leistung in der angenommenen Realsituation (in unserem Fall als Lehrkraft im Fachunterricht) verstanden und die Validität dieser Annahme durch Korrelation mit externen Kriterien geprüft. Die zukünftige Leistung wird demnach nicht direkt beobachtet oder gemessen, sondern näherungsweise durch bereits eingeführte Tests, die ein gleiches oder ähnliches Konstrukt messen, repräsentiert (Kane et al., 1999). In der vorliegenden Validierungsstudie, in der die Annahme validiert werden soll, dass die Ergebnisse des DaZ-Tests als Indikator für ein entsprechend kompetentes Handeln zukünftiger Fachlehrkräfte gelten können, wurden dementsprechend externe Kriterien gewählt, die eine inhaltliche und konzeptuelle Nähe zum Konstrukt *DaZ-Kompetenz* aufweisen und somit einen Nachweis für Konstruktvalidität liefern.

Eine derartige Nähe bietet dabei das Konzept der Lehrerkompetenz von Shulman (1986, 1987). Shulman untergliedert die Wissensbasis der Lehrerkompetenz u. a. in Fachwissen *(content knowledge),* fachdidaktisches Wissen *(pedagogical content knowledge)* und allgemein pädagogisches Wissen *(pedagogical knowledge).* Daran angelehnt wurden bei der Übertragung auf unsere Studie daher linguistisches, mathematikdidaktisches und pädagogisches Wissen als externe Kriterien einbezogen:

- *Linguistisches Wissen* – als konkretisierte Variante des Shulman'schen Fachwissens – weist eine inhaltliche Nähe zur Dimension *Fachregister* des DaZ-Kompetenzmodells (siehe Abb. 1) auf. Als linguistisches Wissen wird im eingesetzten Test das Erkennen, Beurteilen und Einordnen von linguistischen Kategorien verstanden. Diese Anforderung kommt in Auszügen auch in der DaZKom-Teilskala *Fachregister* vor, beispielsweise wenn es um das Erkennen und Kategorisieren sprachlicher Strukturen geht. Die Teilskala *Fachregister* berücksichtigt aber auch Handlungs-

orientierung im Sinne eines auf fachunterrichtliche Situationen bezogenen Problemlöseverhaltens der Lehrkraft.

- *Mathematikdidaktisches Wissen* repräsentiert analog das fachdidaktische Wissen. Da sich das DaZ-Testinstrument des Fachs Mathematik als Referenzdisziplin bedient, wurde das mathematikdidaktische Wissen als zweites externes Kriterium zur Validierung der Testergebnisinterpretation herangezogen.
- *Pädagogisches Wissen* als dritter Bestandteil von Lehrerkompetenz stellt ein Konstrukt dar, das in Bezug auf den Zusammenhang zu den Ergebnissen des DaZ-Tests zu kontrollieren ist. So sind beispielsweise Kenntnisse zum Thema *Umgang mit Heterogenität* Teil des pädagogischen Wissens, aber auch Teil der DaZ-Kompetenz.
- Die Zusammenhänge zwischen den Personenmerkmalen und den Ergebnissen des DaZ-Tests ermöglichen Interpretationen über die Erlernbarkeit von DaZ-Kompetenz. Bezogen auf universitäre Gegebenheiten sind im Speziellen die *Semesterzahl* und die *universitären Lerngelegenheiten* von Interesse, die über die Erlernbarkeit von DaZ-Kompetenz Aufschluss geben können.

4. Forschungsfragen und Hypothesen

Die vorliegende Studie untersucht divergente und konvergente Zusammenhänge der Ergebnisse des im DaZKom-Projekt entwickelten GSL-Testinstruments mit externen Validierungsmerkmalen. Diese Studie geht folgenden Forschungsfragen und Annahmen nach:

1) Inwieweit hängt die DaZ-Kompetenz von Lehramtsstudierenden mit konzeptuell ähnlichen Konstrukten zusammen? Lässt sich DaZ-Kompetenz abgrenzen von

   a) linguistischem Wissen,
   b) pädagogischem Wissen und
   c) mathematikdidaktischem Wissen?

   Im Hinblick auf das linguistische Wissen werden höhere Korrelationen mit der DaZ-Teilskala Fachregister als mit den anderen beiden Teilskalen des DaZ-Tests erwartet. Beim pädagogischen Wissenstest ist ein Zusammenhang mit den Ergebnissen des DaZ-Tests in der Subdimension Umgang mit Heterogenität denkbar, da DaZ-Kompetenz u. a. den Umgang mit sprachlicher Heterogenität betrachtet. Korrelationen mit den anderen Subdimensionen sind eher nicht erwartbar, da der DaZ-Test nicht auf allgemeinpädagogisches Wissen ausgerichtet ist. Ein positiver Zusammenhang zwischen dem DaZ-Test und dem mathematikdidaktischen Wissen ist vorstellbar, weil Mathematik als Bezugsdisziplin fungiert und somit domänenspezifisches Wissen für die Analyse und Förderung von sprachbezogenen Problemsituationen im Mathematikunterricht förderlich sein könnte.

2) In welchem Zusammenhang steht DaZ-Kompetenz mit Merkmalen von Lehramtsstudierenden wie Geschlecht, Studienfach, Semesterzahl und Muttersprache?
   Als Validierungsbeleg kann erwartet werden, dass DaZ-Kompetenz bei Lehramtsstudierenden des Fachs Deutsch höher ausgeprägt ist als bei Lehramtsstudierenden im Fach Mathematik. Zudem ist ein höheres Kompetenzniveau mit steigender Semesterzahl zu vermuten, da davon ausgegangen werden kann, dass mit höherer Semesterzahl die DaZ-Erfahrung (z. B. durch schulische Praktika/universitäre Seminare) ansteigt. Bei den Merkmalen Geschlecht und Erstsprache kann angenommen werden, dass weibliche Studierende und Studierende mit der Erstsprache Deutsch eine höhere DaZ-Kompetenz aufweisen, da den Ergebnissen beispielsweise aus PISA 2012 (OECD, 2014) oder PIAAC 2013 (*Programme for the International Assessment of Adult Competencies;* OECD, 2013b) nach zu urteilen Mädchen und Schülerinnen und Schüler mit deutscher Erstsprache im Durchschnitt höhere Kompetenzen im Bereich Sprache aufweisen.

3) Inwieweit hängt DaZ-Kompetenz mit den universitären Lerngelegenheiten im Bereich DaZ zusammen?
   Hinsichtlich dieser Frage wird angenommen, dass speziell das Absolvieren von Lerngelegenheiten zu Themen wie Zweitspracherwerb, Sprachstandsdiagnostik, grammatische Phänomene des Deutschen oder migrationsbezogene Studieninhalte einen positiven Zusammenhang zu den DaZ-Testergebnissen aufweist.

5. Methode

5.1 Stichprobe

Für die Durchführung der Validierungsstudie wurden $N = 252$ Lehramtsstudierende an fünf Universitäten aus drei Bundesländern (Bayern, Niedersachsen, Nordrhein-Westfalen) getestet. 234 Studierende waren zum Zeitpunkt der Durchführung in ein Lehramtsstudium eingeschrieben. Die belegten Unterrichtsfächer waren hierbei unterschiedlich und breit gestreut (Mathematik, Deutsch, Englisch, Naturwissenschaften, Musik, Geschichte, Kunst, Sport, Sachunterricht), wobei 50 % der Lehramtsstudierenden Mathematik als Unterrichtsfach angaben. 18 der untersuchten Teilnehmenden studierten das Fach Deutsch als Fremdsprache und Germanistik (Master) oder Deutsch als Zweitsprache (Bachelor), da von diesen Studierenden eine besonders hohe Fähigkeit im Bereich DaZ erwartet werden kann. Die Gesamtstichprobe bestand zu 70 % aus Bachelorstudierenden vom ersten bis sechsten Semester und zu 30 % aus Masterstudierenden im ersten bis vierten Semester.

5.2 Instrumente

GSL-Testinstrument

Das Instrument zur Messung von DaZ-Kompetenz besteht aus 68 Items. Jedes dieser Items ist in eine der drei Dimensionen von DaZ-Kompetenz eingeordnet (vgl. Abb. 1). 31 Items werden der Dimension *Fachregister* zugeordnet, 17 Items der Dimension *Mehrsprachigkeit* und 20 Items der Dimension *Didaktik.* Für eine statistische Modellierung wurde keine weitere Differenzierung in die im Kompetenzmodell vorhandenen Subdimensionen vorgenommen, da die Anzahl der Items pro Subdimension zu gering gewesen wäre.

Der Test besteht aus 16 Aufgabenunits, die jeweils aus zwei bis neun Einzelitems bestehen. Jede Aufgabenunit beginnt mit einem authentischen Stimulus, der entweder ein Fallbeispiel, eine Lehrer-Schüler-Interaktion, ein schriftliches Schülerprodukt oder eine Mathematiktextaufgabe mit potenziellen sprachlichen Schwierigkeiten beinhaltet. Der Test umfasst die gängigen Antwortformate (Bortz & Döring, 2006): 32 Aufgaben mit einem geschlossenen Antwortformat, 14 Aufgaben mit halboffenem und 22 Aufgaben mit offenem Antwortformat. Die in Abbildung 2 dargestellte Beispielaufgabe zeigt eine Mathematiktextaufgabe mit Wortschwierigkeiten und wird der Dimension *Fachregister* zugeordnet.[3]

Die Testdaten wurden auf Basis der Item-Response-Theorie (IRT) Rasch-skaliert (vgl. Carlson et al., in Vorbereitung). Die Item-Fit-Werte zeigen, dass die Testitems gut mit dem Raschmodell übereinstimmen. Auch die Itemtrennschärfen liefern akzeptable Werte (M = 0.32, SD = 0.10, MIN = 0.13, MAX = 0.57). Zudem zeigen die Werte zur Itemschwierigkeit (M = 0.07, SD = 0.88) und zur Personenfähigkeit (M = 0.00, SD = 0.78) eine hohe Überschneidung. Die *EAP-Reliabilität* = 0.80 des DaZ-Tests liegt in einem annehmbaren Bereich, ebenso die Reliabilitäten der Subdimensionen (1 = Fachregister (α = 0.74), 2 = Didaktik (α = 0.69), 3 = Mehrsprachigkeit (α = 0.66)). Zudem zeigen die Korrelationen der Dimensionen untereinander zu erwartende Werte ($r_{1,2}$ = .75, $r_{2,3}$ = .28, $r_{1,3}$ = .62). Die Analyse der Dimensionalität hat gezeigt, dass das ein- und das dreidimensionale Modell sehr ähnliche Informationskriterien liefern (eindimensionales Modell: AIC = 12716, BIC = 12960, CAIC = 13029; dreidimensionales Modell: AIC = 12604, BIC = 12866, CAIC = 12940). Im Sinne des Parsimonitätsprinzips wird DaZ-Kompetenz als eindimensionales Konstrukt verstanden. Im Folgenden werden wir neben den Testwerten aus der eindimensionalen Skalierung auch die Scores aus den drei Subdimensionen auswerten, um hier differenziellen Hypothesen nachgehen zu können.

3 Das vorliegende Item wurde wegen schlechter psychometrischer Werte verworfen. Es dient an dieser Stelle jedoch als Beispiel für den Aufbau einer Aufgabenunit und verdeutlicht durch die Wortwahl, dass die Testaufgaben nicht für Experten der Germanistik bzw. Linguistik mit deren Terminologien entwickelt wurden.

Peter möchte in der Pause im Schulkiosk Süßigkeiten kaufen. Er kauft 10 Bonbons für jeweils 20 Cent. Sein Freund Max kann nicht widerstehen und kauft sich ebenfalls 5 Leckereien für je 50 Cent. Wer hat mehr Geld ausgegeben – Peter oder Max?

1. Nennen Sie vier Wörter aus dem Aufgabenbeispiel, die für einen DaZ-Lernenden schwer zu verstehen sein könnten.

2. Bei welchen sprachlichen Referenzen im Text, die für die Beantwortung der Aufgabe relevant sind, könnten DaZ-Lernende Schwierigkeiten haben? *Erläutern Sie jeweils die Schwierigkeit.*

Quelle: Bezirksregierung Münster (2008): Sprachförderung als Aufgabe aller Fächer – Mathematik – Gesamtschulen, S. 13

Abb. 2: DaZKom-Aufgabenbeispiel zur Dimension Fachregister

Hintergrundmerkmale und Lerngelegenheiten

Neben dem DaZ-Test wurde ein Fragebogen zu Hintergrundmerkmalen und zu Lerngelegenheiten im Bereich DaZ eingesetzt. Der Fragebogen zu den Hintergrundmerkmalen beinhaltet Fragen zum Geschlecht, zur Muttersprache, zur Anzahl der studierten Semester sowie zum Studienfach. Die Skala zu Lerngelegenheiten im Bereich DaZ besteht aus 13 dichotomen Items, die nach dem Vorkommen DaZ-spezifischer Themen im Lehramtsstudium fragen. Angelehnt an TEDS-M soll der Fragebogen zu den Lerngelegenheiten die Variation des Wissens der Studierenden und der Lerngelegenheiten selbst dokumentieren (Floden, 2002).

Linguistischer Wissenstest

Der linguistische Wissenstest stammt aus dem *LiKoM*-Teilprojekt *Sprachreflexion* (Nottbusch, Sahel, Civak, Stanojević & Wiejowski, 2014). Der genannte Test diente in diesem Projekt der Ermittlung von Sprachkompetenz hinsichtlich der Erkennung, Beurteilung und Einordnung sprachlicher Phänomene (Civak, Stanojević, Stummeier & Vogel, 2012). Der Test besteht aus 85 dichotomen Items, die die Hauptbereiche der allgemeinen Linguistik repräsentieren. Je fünf Items gehören zu einer Frage. Dabei geht es z. B. darum, aus fünf möglichen Wörtern zwei zu identifizieren, die dieselbe Silbenstruktur aufweisen. Der Test weist über die Gesamtskala einen EAP-Reliabilitätskoeffizienten von 0.92 auf.

Mathematikdidaktischer Wissenstest

Das mathematikdidaktische Testinstrument basiert auf dem TEDS-M-Instrument für mathematikdidaktisches Wissen und stellt eine gekürzte Fassung mit 31 Items (29 geschlossene und 2 offene Antwortformate) dar (TEDS-shortM; Buchholtz et al., 2012). Ziel dieses Fragebogens ist es, mathematikdidaktisches Wissen in der Lehrerausbildung zu erfassen. Die Testaufgaben bilden zwei allgemeine Gegenstandsbereiche der Mathe-

matikdidaktik ab. Im Bereich der *Stoffdidaktik* wird u. a. die Reflexion von Alltags- und Fachsprache bei mathematischer Begriffsbildung behandelt. Der Bereich *Unterrichtsdidaktik* beschäftigt sich u. a. mit der Kenntnis von Heterogenität im Mathematikunterricht. Der Gesamttest zeigte einen EAP-Reliabilitätskoeffizienten von 0.69.

Pädagogischer Wissenstest
Der eingesetzte Test wurde ebenfalls im Rahmen der TEDS-M-Studie entwickelt und eingesetzt. Der Test dient der Kompetenzmessung pädagogischen Wissens, das als kognitive Komponente professioneller Kompetenz angehender Lehrerinnen und Lehrer verstanden wird (König & Blömeke, 2010). Die Konstruktion dieses Tests basiert auf den beruflichen Anforderungen von Lehrpersonen (siehe z. B. Standards für die Lehrerbildung; Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004). Diese Anforderungen werden in vier Teilskalen dargestellt: *Umgang mit Heterogenität, Strukturierung, Klassenführung/Motivierung* und *Leistungsbeurteilung.* Die Kurzfassung des Testinstruments besteht aus 18 Testaufgaben (10 geschlossene und 8 offene Antwortformate). Der Test folgt der Fragestellung, inwieweit fachübergreifende, erziehungs- bzw. bildungswissenschaftliche, pädagogische Kompetenzen in der Lehrerausbildung zu finden sind. Die Reliabilität zeigt für die Gesamtskala einen Kennwert von 0.65. Für die Subskalen *Umgang mit Heterogenität* ($\alpha = 0.69$), *Strukturierung* ($\alpha = 0.64$), *Klassenführung/Motivierung* ($\alpha = 0.66$) und *Leistungsbeurteilung* ($\alpha = 0.44$) zeigen sich für fast alle Skalen akzeptable Reliabilitäten.

5.3 Vorgehen der Datenerhebung

Die Datenerhebung erfolgte im Winter 2013. Der Test wurde in eigenen Testsitzungen von geschulten Mitarbeiterinnen und Mitarbeitern an den jeweiligen Universitäten durchgeführt. Die Durchführungsdauer der Tests und Fragebögen betrug für den DaZ-Test 60 Minuten, für die Lerngelegenheiten 15 Minuten, für den pädagogischen Wissenstest bzw. den mathematikdidaktischen Wissenstest 25 Minuten und für den linguistischen Wissenstest 15 Minuten. Die Teilnahme war freiwillig, anonym und wurde mit 10 € pro Stunde vergütet. Der DaZ-Test wurde in vier Testheftversionen eingesetzt, die sich sowohl teilweise in den Aufgaben als auch in der Reihenfolge unterschieden. Alle Testheftversionen verfügten über drei Aufgabenblöcke, die in allen Versionen vorhanden waren und an derselben Stelle im Testheft angeordnet waren. Alle Teilnehmenden bearbeiteten eine Version des DaZ-Tests, den linguistischen Wissenstest und den Fragebogen zu den Lerngelegenheiten. Um die Testzeit möglichst kurz zu halten, wurde jeweils der Hälfte der Teilnehmerinnen und Teilnehmer entweder der mathematikdidaktische oder der pädagogische Wissenstest vorgelegt.

5.4 Statistisches Vorgehen

Die Item- und Skalenanalysen für die eingesetzten Tests wurden auf der Basis des Rasch-Modells (z. B. Rost, 2004) mit dem Programm *ConQuest* (Adams, Wu & Wilson, 2012) durchgeführt. Die Personenmesswerte wurden anhand von Personenfähigkeitsschätzern, Weighted Likelihood Estimates (WLE), bestimmt. Diese beruhen ausschließlich auf den Antworten der Probanden auf die Testaufgaben. Wegen ihrer geringen durchschnittlichen Abweichung vom wahren Kompetenzwert eignen sich WLE-Schätzer besonders gut zur Bestimmung individueller Kompetenzausprägungen (Rost, 2004). Zusammenhänge mit externen Maßen wurden anhand von bivariaten Korrelationen berechnet. Die interne Konsistenz der Skalen wurde durch EAP-Reliabilitäten geprüft (z. B. Wilson, 2005). Fehlende Antworten, d. h. ungültige und nicht ausgefüllte Antwortfelder, wurden als falsch bewertet. Für den mathematikdidaktischen sowie den pädagogischen Wissenstest lagen aufgrund des Testdesigns nur für die Hälfte der Stichprobe Daten vor.

6. Ergebnisse

6.1 Zusammenhänge von DaZ-Kompetenz mit linguistischem, pädagogischem und mathematikdidaktischem Wissen

Tabelle 1 zeigt die bivariaten Korrelationen zwischen den Ergebnissen des Gesamttests bzw. für die drei Teilskalen des DaZ-Tests jeweils mit den Ergebnissen der zusätzlich eingesetzten Wissenstests. Anhand der Resultate wird sichtbar, dass der linguistische Wissenstest sowohl mit der Gesamtskala des DaZ-Tests (r = .25) als auch mit den Teilskalen *Fachregister* (r = .23) und *Didaktik* (r = .19) des DaZ-Tests signifikant positiv korreliert. Die Ergebnisse des pädagogischen Wissenstests in der Dimension *Umgang mit Heterogenität* korrelieren signifikant sowohl mit der Gesamtskala (r = .21) als auch mit der Teilskala *Mehrsprachigkeit* (r = .23). Die anderen Dimensionen des pädagogischen Wissenstests korrelieren nicht signifikant mit dem DaZ-Test. Die Ergebnisse des mathematikdidaktischen Wissenstests stehen weder mit der DaZ-Gesamtskala noch mit den Teilskalen in Zusammenhang. Demnach grenzt sich DaZ-Kompetenz deutlich von mathematikdidaktischem Wissen ab. Insgesamt fällt die Höhe der Korrelationskoeffizienten sehr gering aus. Alle Koeffizienten sind kleiner als r = .30. Nach Bortz (2005) kann der Zusammenhang zwischen DaZ-Kompetenz und dem linguistischen sowie dem pädagogischen Wissen als gering bezeichnet werden.

	Gesamtskala	Teilskalen		
		Fachregister	Didaktik	Mehrsprachigkeit
Linguistisches Wissen	.25**	.23**	.19**	.09
Pädagogisches Wissen				
Umgang mit Heterogenität	.21*	.14	.11	.23*
Unterrichtsstrukturierung	.09	.07	.11	.05
Klassenführung/Motivation	.09	.05	−.09	.18
Leistungsbeurteilung	.08	.09	.01	.03
Mathematikdidaktisches Wissen				
Stoffdidaktik	.11	.14	−.01	.02
Unterrichtsdidaktik	.18	.17	.00	.11

** $p < 0.01$; * $p < 0.05$

Tab. 1: Korrelation zwischen DaZ-Kompetenz (Gesamtskala und Teilskalen) mit dem linguistischen, mathematikdidaktischen und pädagogischen Wissen

6.2 Zusammenhänge mit Studienfach, Semesterzahl, Geschlecht und Erstsprache

Tabelle 2 zeigt die Zusammenhänge zwischen den Ergebnissen der ein- und der dreidimensionalen DaZ-Kompetenzskalen und den Personenmerkmalen Geschlecht, Studienfach, Semesterzahl und Muttersprache. Lehramtsstudierende mit dem Unterrichtsfach *Deutsch* erreichen eine höhere DaZ-Kompetenz als Studierende anderer Studienfächer ($r = .14$). Das Studienfach *Mathematik* weist keine signifikanten Korrelationen mit den Testergebnissen des DaZ-Tests auf. Dies deutet darauf hin, dass Lehramtsstudierende des Unterrichtsfachs *Mathematik* keine besseren Ergebnisse im DaZ-Test erzielen als Studierende anderer Unterrichtsfächer. Die *Semesterzahl* zeigt jedoch einen deutlichen Zusammenhang zu den Testergebnissen des eindimensionalen Modells ($r = .17$) und der Teilskala *Mehrsprachigkeit* des dreidimensionalen Modells ($r = .19$). Ein fortgeschrittenes Semester führt insbesondere durch steigende Kompetenz im Bereich *Mehrsprachigkeit* zu besseren Testergebnissen. Beim Merkmal *Geschlecht* zeigt sich ein signifikanter Zusammenhang zum eindimensionalen Modell ($r = -.17$). Teilnehmerinnen erzielen demnach bessere Ergebnisse als Teilnehmer. Im dreidimensionalen Modell zeigt sich für die Teilskala *Fachregister* ein signifikanter Zusammenhang mit der *Muttersprache Deutsch* ($r = -.24$). Demzufolge erzielen Teilnehmerinnen und Teilnehmer mit Deutsch als Erstsprache bessere Ergebnisse für die Teilskala *Fachregister* als Nicht-Erstsprachlerinnen und Nicht-Erstsprachler.

	Gesamtskala	Teilskalen		
		Fachregister	Didaktik	Mehrsprachigkeit
Fach Mathematik	−.03	.03	−.05	−.09
Fach Deutsch	.14*	.14	.08	.09
Semesterzahl	.17*	.09	.12	.19**
Geschlecht (w = 0; m = 1)	−.17*	−.11	−.03	−.09
Erstsprache (Deutsch = 0; andere = 1)	−.13	−.24**	.03	.01

** p < 0.01; * p < 0.05

Tab. 2: Korrelationen der Ergebnisse des DaZ-Tests mit Personenmerkmalen

6.3 Zusammenhänge mit universitären Lerngelegenheiten im Bereich DaZ

In Tabelle 3 sind die Korrelationen zwischen den Ergebnissen des DaZ-Tests und der universitären Lerngelegenheiten dargestellt. Die Items *Teilgebiete der Linguistik* (r = .17), *Phänomene des Zweitspracherwerbs* (r = .19), *Unterschiede zwischen Fremd- und Zweitsprache* (r = .18), *Sprachstandsdiagnostik* (r = .19) sowie *Unterstützung des sprachlichen Lernprozesses durch Scaffolding* (r = .22) zeigen signifikante Zusammenhänge mit der Teilskala *Mehrsprachigkeit.* Das Item *Unterschiede zwischen mündlich und schriftlich geprägter Sprache* zeigt einen hohen Zusammenhang mit den Teilskalen *Fachregister* (r = .18) und *Mehrsprachigkeit* (r = .23). Lehramtsstudierende, die viele universitäre Lerngelegenheiten im Bereich DaZ wahrgenommen haben, erreichen demnach eine höhere DaZ-Kompetenz als Studierende, die wenige Lerngelegenheiten in diesem Gebiet genutzt haben.

Um abzuschätzen, inwieweit den untersuchten Personenmerkmalen und den verschiedenen Lerngelegenheiten jeweils ein spezifischer Vorhersagebeitrag auf die DaZ-Kompetenz zukommt, haben wir eine Regressionsanalyse gerechnet, in der alle Merkmale gleichzeitig einbezogen wurden. Bei Kontrolle aller übrigen Merkmale weisen die Prädiktoren Semesterzahl (β = 0.19), Erstsprache (β = 0.18) und Studienfach Deutsch (β = 0.19) einen signifikanten Vorhersagebeitrag auf.

Mit längerer Studiendauer steigt demnach die DaZ-Kompetenz. Sie fällt außerdem höher bei Studierenden mit dem Studienfach Deutsch aus und wenn die Erstsprache Deutsch ist. Durch die Analyse können insgesamt 14 Prozent der Varianz in der DaZ-Kompetenz aufgeklärt werden.

	Teilskalen		
	Fachregister	**Didaktik**	**Mehrsprachigkeit**
(Teil-)Gebiete der Linguistik	.07	−.05	.17*
Grammatik des Deutschen	−.03	.06	.05
Unterschiede zwischen mündlich und schriftlich geprägter Sprache	.18*	.00	.23**
Erwerb von Bildungssprache	.04	−.04	.04
Phänomene des Zweitspracherwerbs	.02	−.09	.19**
Erwerbsequenzen sprachlicher Entwicklung	.09	.01	.11
Unterschiede zwischen Fremd- und Zweitspracherwerb	.04	−.09	.18*
Migration und Mehrsprachigkeit	.07	−.02	.15*
Sprachliche Vielfalt in der Schule	−.02	−.01	.08
Sprachstandsdiagnostik	.06	.06	.19**
Sprachförderung	.08	−.11	.10
Unterstützung des sprachlichen Lernprozesses durch Scaffolding	.01	−.02	.22**
Sprachsysteme von Zuwanderungssprachen	.03	−.05	.06

** $p < 0.01$; * $p < 0.05$

Tab. 3: Korrelationen der Ergebnisse des DaZ-Tests mit universitären Lerngelegenheiten

7. Diskussion

Das Ziel dieser Studie bestand darin, die Validität des im Rahmen des BMBF-Projektes DaZKom entwickelten Testinstruments zu untersuchen. Mit dem GSL-Testinstrument sollen Kompetenzen von Lehramtsstudierenden im Bereich Deutsch als Zweitsprache gemessen werden. In Anlehnung an den argumentationsbasierten Ansatz nach Kane (2013) wurde überprüft, inwieweit die Testscores aus dem DaZ-Test im Zusammenhang mit Merkmalen stehen, die auf die Plausibilität der Testwertinterpretation schließen lassen.

Zuerst wurde geprüft, ob und inwieweit DaZ-Kompetenz von Lehramtsstudierenden mit konzeptuell ähnlichen Konstrukten wie linguistischem Fachwissen, mathematikdidaktischem und pädagogisch-unterrichtsbezogenem Wissen zusammenhängt. Um diese Zusammenhänge differenziert untersuchen zu können, wurden neben dem DaZ-Gesamttest auch die Scores aus den drei DaZ-Teilskalen *(Fachregister, Mehrsprachigkeit, Didaktik)* herangezogen.

Im Speziellen wurde ein Zusammenhang zwischen der Teilskala *Fachregister* und dem linguistischen Fachwissen erwartet. Beide Skalen weisen Überschneidungen im

Bereich des Erkennens und des Kategorisierens sprachlicher Strukturen auf. Der prognostizierte signifikante Zusammenhang konnte hier durch die Daten bestätigt werden. Die niedrigen Korrelationskoeffizienten weisen darauf hin, dass es sich nur um eine begrenzte Übereinstimmung der von den Instrumenten erfassten Kompetenzdimensionen und -facetten handelt.

Die Teilskala *Fachregister* des GSL-Tests geht im Sinne der Kompetenzmodellierung weit über das isolierte Erkennen und Kategorisieren sprachlicher Strukturen, das in einem Linguistiktest verlangt wird, hinaus. Sprachliche Strukturen müssen in der Teilskala *Fachregister* von vornherein im Kontext der in den Items jeweils dargestellten Situation hinsichtlich ihrer Funktionalität erkannt und kategorisiert sowie darüber hinaus mit Blick auf den abgebildeten Problemzusammenhang fachlichen Lernens bewertet werden. Die geringe Korrelation unterstreicht daher, dass ein linguistischer Wissenstest zwar notwendiges Wissen über sprachliche Strukturen erfasst, aber die in der Subdimension *Fachregister* des DaZ-Tests operationalisierten Formen komplexen Problemlösens mit Fokus auf Sprache nicht hinreichend abbilden kann.

Darüber hinaus zeigte sich ein Zusammenhang der DaZ-Subdimension *Didaktik* mit dem linguistischen Fachwissen. Dieses Ergebnis lässt sich durch die Struktur der Items erklären. Testaufgaben, die aufgrund ihres Sprachförderungs- bzw. Diagnosefokus der Dimension *Didaktik* zugeordnet wurden, greifen auf inhaltlicher Ebene häufig linguistische Phänomene auf.

Der Zusammenhang zwischen den Scores aus dem GSL-Testinstrument und den Subskalen aus dem pädagogischen Wissenstest fiel ebenfalls wie erwartet positiv aus. Es zeigt sich ein Zusammenhang der Dimension *Umgang mit Heterogenität* des pädagogischen Wissenstests mit der Teilskala *Mehrsprachigkeit* des DaZ-Tests. Da unter *Umgang mit Heterogenität* u. a. die Berücksichtigung ethnischer und leistungsmäßiger Differenzierung zu verstehen ist (Blömeke et al., 2010), erscheint ein Zusammenhang mit der Dimension *Mehrsprachigkeit* an dieser Stelle sehr schlüssig. Die Bereiche *Unterrichtsstrukturierung, Klassenführung/Motivierung* sowie *Leistungsbeurteilung* des pädagogischen Wissenstests zeigen keinen Zusammenhang mit dem GSL-Test. Da dieser Test DaZ-spezifisches Wissen fokussiert und der pädagogische Wissenstest eher allgemeindidaktisches Wissen abfragt, waren in diesen Bereichen auch keine Zusammenhänge zu erwarten.

Hinsichtlich des Zusammenhangs zwischen DaZ-Kompetenz und mathematikdidaktischem Wissen zeigen die Ergebnisse, dass es entgegen der vorab formulierten Erwartungen keinen Zusammenhang gibt. Da der DaZ-Test das Schulfach Mathematik als Bezugsdisziplin verwendet, konnte angenommen werden, dass Studierende mit hohem mathematikdidaktischem Wissen möglicherweise über eine höhere Sensibilität für mathematikspezifische Fachsprache verfügen. Die Resultate bestätigen diese Vermutung jedoch nicht. Damit wird vor allem klar, dass fachsprachliche Kompetenz nicht ohne Weiteres mit dem im DaZ-Test abgefragten Wissen über bildungssprachliche Anforderungen des Mathematikunterrichts gleichgesetzt werden kann. Grundsätzlich kann bemerkt werden, dass sich die DaZ-Kompetenz im Fach Mathematik vom mathematikdidaktischen Wissen deutlich abgrenzt.

Zusammenfassend lässt sich deshalb festhalten: Es bestehen plausible Zusammenhänge zwischen einzelnen Subskalen des GSL-Testinstruments und denen des linguistischen und pädagogischen Wissenstests. Absolut betrachtet, fallen jedoch die Korrelationskoeffizienten in ihrer Höhe eher niedrig aus.

In einem zweiten Schritt wurde geprüft, ob die Ergebnisse des DaZ-Tests in einem plausiblen Zusammenhang mit Hintergrundmerkmalen der Lehramtsstudierenden stehen. Es wurde vermutet, dass Lehramtsstudentinnen eventuell aufgrund von besseren Sprachkompetenzen (vgl. OECD, 2013a, 2013b, 2014) sensibler für DaZ-spezifische Lernsituationen sind als männliche Studenten. Diese Vermutung konnte bestätigt werden, wenngleich der Unterschied – absolut betrachtet – wiederum nicht besonders stark ausfällt. Zudem zeigte die studienfachbezogene Analyse, dass Studierende des Deutschlehramts bessere Ergebnisse im DaZ-Test erreichten als Studierende anderer Studienfächer. Dies erscheint vor dem Hintergrund, dass DaZ-Kompetenz u. a. sprachliches Wissen und die Fähigkeit zur Reflexion über Sprache beinhaltet, plausibel. Analog zum Befund hinsichtlich des mathematikdidaktischen Wissens zeigte sich keine höhere DaZ-Kompetenz bei Studierenden mit dem Unterrichtsfach Mathematik, was in Anbetracht der Bezugsdisziplin Mathematik des GSL-Testinstruments eigentlich erwartet wurde. Es konnte darüber hinaus gezeigt werden, dass mit steigender Semesterzahl die DaZ-Kompetenz in unserer Stichprobe ansteigt. Dies ist ein Indiz dafür, dass Lerngelegenheiten eine große Rolle beim Erlernen von DaZ-Kompetenz spielen. Es wird dabei davon ausgegangen, dass bei steigender Semesterzahl eine erhöhte Wahrscheinlichkeit besteht, dass Lerngelegenheiten wahrgenommen werden.

Ein interessanter Befund zeigte sich hinsichtlich des Vergleichs der Studierenden, die Deutsch selbst als Zweitsprache gelernt haben, mit den Studierenden mit Deutsch als Erstsprache: Denkbar wäre an dieser Stelle gewesen, dass Studierende mit Deutsch als Zweitsprache aufgrund der eigenen Spracherfahrung sensibler für sprachbezogene Lernsituationen sind. Diese Vermutung konnte jedoch nicht bestätigt werden. Deutsch-Erstsprachlerinnen und -Erstsprachler erreichten entgegen der aufgestellten Vermutung in der Subdimension *Fachregister* höhere Testwerte als Nicht-Deutsch-Erstsprachlerinnen und -Erstsprachler.

In einem dritten Schritt wurde der Zusammenhang zwischen der Nutzung universitärer Lerngelegenheiten im Bereich Deutsch als Zweitsprache und den Ergebnissen des DaZ-Tests untersucht. Es wurde erwartet, dass Studierende höhere DaZ-Testergebnisse erzielen, je mehr DaZ-bezogene Lerngelegenheiten von ihnen wahrgenommen wurden. Diese Annahme konnte bestätigt werden. Es stellte sich zudem heraus, dass die genutzten Lerngelegenheiten in besonderem Zusammenhang zu der DaZ-Teilskala *Mehrsprachigkeit* stehen. Möglicherweise liegt hier ein curricularer Schwerpunkt der universitären Lehrerausbildung allgemein bzw. der DaZ-Module im Besonderen.

Hier liegen die Grenzen der eingesetzten Fragebogenskala. Um genauer zu klären, welche Lerngelegenheiten in welchem Umfang prädiktiv für den Erwerb von DaZ-Kompetenz sind, müssten diese deutlich differenzierter und bestenfalls in einem längsschnittlichen Design erhoben werden. Im Rahmen dieser Validierungsstudie war dies jedoch aufgrund der begrenzten Testzeit nicht möglich. Eine weitere Einschränkung

dieser Studie liegt in der Begrenztheit der Stichprobe. Aufgrund der heterogenen universitären Lerngelegenheiten im Lehramtsstudium im Bereich DaZ sollten weitere Universitätsstandorte einbezogen werden. Zukünftig sollten daher Replikationsstudien durchgeführt werden. Ein weiteres Forschungsdesiderat ergibt sich aus dem Befund, dass nur Lehramtsstudierende mit dem Unterrichtsfach Deutsch eine höhere DaZ-Kompetenz aufweisen als andere Studierende. Auch dieses Ergebnis sollte anhand anderer Stichproben repliziert werden. Inhaltlich wirft dies die Forschungs- und Entwicklungsfrage auf, inwieweit und welche Lernarrangements hier für die fachdidaktische Lehramtsausbildung an Universitäten konzipiert werden könnten. Dafür müssten gezielt Fortbildungsveranstaltungen entwickelt werden. Der in dieser Arbeit vorgestellte DaZ-Kompetenztest könnte dabei als Instrument für die Evaluation eingesetzt werden. Ein zweiter Entwicklungsschritt, der noch aussteht, ist die normative Bewertung von DaZ-Kompetenz. Welches Niveau von DaZ-Kompetenz sollte bei angehenden Lehrkräften als Mindeststandard, welches als Regel- und Optimalstandard gelten? Diese Grenzen müssten im Rahmen eines Standardsettingverfahrens (Tiffin-Richards & Köller, 2010), an dem Expertinnen und Experten aus unterschiedlichen Bereichen zu beteiligen sind, diskutiert und festgelegt werden.

Literatur

Adams, R. J., Wu, M., & Wilson, M. (2012). *ACER ConQuest* [Computer Software]. Australian Council for Educational Research (ACER). http://www.acer.edu.au/conquest/overview2 [29. 05. 2014].

Becker-Mrotzek, M., Hentschel, B., Hippmann, K., & Linnemann, M. (2012). *Sprachförderung in deutschen Schulen – die Sicht der Lehrerinnen und Lehrer. Ergebnisse einer Umfrage unter Lehrerinnen und Lehrern*. Durchgeführt von IPSOS (Hamburg) im Auftrag des Mercator-Instituts für Sprachförderung und Deutsch als Zweitsprache. Köln: Universität.

Bezirksregierung Münster (2008). *Sprachförderung als Aufgabe aller Fächer – Mathematik – Gesamtschulen.* Münster: Bezirksregierung.

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

Blossfeld, H.-P., Bos, W., Lenzen, D., Müller-Böling, D., Oelkers, J., Prenzel, M., & Wößmann, L. (2007). *Bildungsgerechtigkeit. Jahresgutachten 2007*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bortz, J. (2005). *Statistik: Für Human- und Sozialwissenschaftler* (6., vollst. überarb. u. aktual. Aufl.). Heidelberg: Springer.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* (4., überarb. Aufl.). Heidelberg: Springer.

Buchholtz, N., Scheiner, T., Döhrmann, M., Suhl, U., Kaiser, G., & Blömeke, S. (2012). *TEDS-shortM: Kurzfassung der mathematischen und mathematikdidaktischen Testinstrumente aus TEDS-M, TEDS-LT und TEDS-Telekom*. Hamburg: Universität.

Carlson, S. A., Hammer, S., Rosenbrock, S., Köker, A., Ehmke, T., Koch-Priewe, B., & Ohm, U. (in Vorbereitung). *Test Instrument for the Assessment of Pre-Service Teachers' Teaching Competencies in the Field of German as a Second Language*.

Civak, S., Stanojević, M. M., Stummeier, C., & Vogel, R. (2012). Kompetenzentwicklung bei Germanistik- und Physikstudierenden – Trainingseffekt durch das Studium? In U. Preußer

& N. Sennewald (Hrsg.), *Literale Kompetenzentwicklung an der Hochschule* (S. 325–346). Frankfurt a. M.: Peter Lang.

Commins, N. L., & Miramontes, O. B. (2006). Addressing linguistic diversity from the outset. *Journal of Teacher Education, 57*(3), 240–246.

de Jong, E. J. (2013). Preparing Mainstream Teachers for Multilingual Classrooms. *Association of Mexican-American Educators (AMAE) Special Invited Issue, 7*(2), 40–49.

de Jong, E. J., & Harper, C. A. (2005). Preparing Mainstream Teachers for English-Language Learners: Is Being a Good Teacher Good Enough? *Teacher Education Quarterly, 32*(2), 101–124.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine. The power of human intuition and expertise in the era of the computer*. Oxford: Basil Blackwell.

Echevarria, J., Short, D., & Vogt, M. (2008). *Implementing the SIOP® Model through Effective Professional Development and Coaching*. Boston: Allyn & Bacon.

Floden, R. (2002). The measurement of opportunity to learn. In A. C. Porter & A. Gamoran (Hrsg.), *Methodological advances in cross-national surveys of educational achievement* (S. 231–266). Washington, D. C.: National Academy Press.

Gibbons, P. (2002). *Scaffolding language, scaffolding learning: teaching second language learner in the mainstream classroom*. Portsmouth: Heinemann.

Gogolin, I. (2009). Zweisprachigkeit und die Entwicklung bildungssprachlicher Fähigkeiten. In I. Gogolin & U. Neumann (Hrsg.), *Streitfall Zweisprachigkeit – The Bilingualism Controversy* (S. 263–280). Wiesbaden: VS Verlag für Sozialwissenschaften.

Halliday, M. A. K. (1978). *Language as Social Semiotic*. London: Edward Arnold.

Hopp, H., Thoma, D., & Tracy, R. (2010). Sprachförderkompetenz pädagogischer Fachkräfte: Ein sprachwissenschaftliches Modell. *Zeitschrift für Erziehungswissenschaft, 13*(4), 609–629.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Kecker, G. (2011). *Validierung von Sprachprüfungen. Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen*. Frankfurt a. M.: Lang.

Koch, P., & Oesterreicher, W. (1994). Schriftlichkeit und Sprache. In H. Günther & O. Ludwig (Hrsg.), *Schrift und Schriftlichkeit: Ein interdisziplinäres Handbuch internationaler Forschung* (1. Halbbd., S. 587–604). Berlin/New York: de Gruyter.

Köker, A., Rosenbrock, S., Ohm, U., Carlson, S. A., Ehmke, T., Hammer, S., & Koch-Priewe, B. (2015). DaZKom – Ein Modell von Lehrerkompetenz im Bereich Deutsch als Zweitsprache. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Hrsg.), *Welche Kompetenzen brauchen Lehramtsstudierende und angehende ErzieherInnen? Theoretische und empirische Zugänge*. Bad Heilbrunn: Klinkhardt.

Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004). *Standards für die Lehrerbildung: Bildungswissenschaften.* http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards_Lehrerbildung-Bericht_der_AG.pdf [29. 05. 2014].

König, J., & Blömeke, S. (2010). *Pädagogisches Unterrichtswissen (PUW). Dokumentation der Kurzfassung des TEDS-M-Testinstruments zur Kompetenzmessung in der ersten Phase der Lehrerausbildung*. Berlin: Humboldt-Universität.

Leung, C. (2007). English as an additional language policy: issues of inclusive access and language learning in the mainstream. *NALDIC Quarterly, 3*(1), 16–26.

Li, X., & Zhang, M. (2004). Why Mei still cannot read and what can be done. *Journal of Adolescent & Adult Literacy, 48*(2), 92–101.

Lucas, T., & Grinberg, J. (2008). Responding to the linguistic reality of mainstream classrooms. Preparing all teachers to teach English language learners. In M. Cochran-Smith, S. Feiman-

Nemser & J. D. McIntyre (Hrsg.), *Handbook of Research on Teacher Education. Enduring Questions and Changing Contexts* (3. Aufl., S. 606–636). New York: Routledge.

Maas, U. (2008). *Sprache und Sprachen in der Migrationsgesellschaft*. Göttingen: V&R unipress mit Universitätsverlag Osnabrück.

Nottbusch, G., Sahel, S., Civak, S., Stanojević, M., & Wiejowski, S. (2014*). LiKoM – Teilprojekt „Entwicklung sprachreflexiver Kompetenzen" – Sprachtest*. Sprachkompetenztest. http://www.uni-bielefeld.de/lili/projekte/likom/Ergebnisse.html [24. 03. 2014].

OECD (2013a). *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II)*. PISA, OECD Publishing.

OECD (2013b). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. OECD Publishing.

OECD (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I)*. PISA, OECD Publishing.

Ohm, U. (2009). Zur Professionalisierung von Lehrkräften im Bereich Deutsch als Zweitsprache: Überlegungen zu zentralen Kompetenzbereichen für die Lehrerausbildung. *Zeitschrift für Interkulturellen Fremdsprachenunterricht, 14*(2), 28–36.

Ohm, U. (2010). Fachliche Schwierigkeiten sind sprachliche Schwierigkeiten. Müssen Fachlehrer und Ausbilder auch Sprachlehrer sein? In C. Chlosta & M. Jung (Hrsg.), *DaF integriert: Literatur – Medien – Ausbildung. Tagungsband der 36. Jahrestagung des Fachverbandes Deutsch als Fremdsprache 2008* (S. 271–284). Göttingen: Universitätsverlag.

Prediger, S., & Özdil, E. (Hrsg.) (2011). *Mathematiklernen unter Bedingungen der Mehrsprachigkeit*. Münster: Waxmann.

Rost, J. (2004). *Testtheorie Testkonstruktion* (2., vollst. überarb. u. erw. Aufl.). Bern: Huber.

Scarcella, R. (2003). *Academic English. A conceptual framework*. Linguistic minority research institute, University of California.

Schleppegrell, M. J. (2004). *The language of school: A functional linguistics perspective*. Mahwah: Lawrence Erlbaum Associates.

Schleppegrell, M. J. (2007). The Linguistic Challenges of Mathematics. Teaching and Learning: a Research Review. *Reading & Writing Quarterly, 23*, 139–159.

Schmidt, W. H., Tatto, M. T., Bankov, K., Blömeke, S., Cedillo, T., Cogan, L., Han, S. I., Houang, R., Hsieh, F. J., Paine, L., Santillan, M., & Schwille, J. (2007). *The Preparation Gap: Teacher Education for Middle School Mathematics in Six Countries. MT21 Report.* East Lansing: Michigan State University.

Shanahan, T., & Shanahan, C. (2008). Teaching Disciplinary Literacy to Adolescents: Rethinking Content-Area Literacy. *Harvard Educational Review, 78*(1), 40–59.

Shulman, L. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*(2), 4–14.

Shulman, L. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review,* 57(1), 1–21.

Simmons, R. D. (2009). The efficacy of Florida's approach to in-service English speakers of other languages (ESOL) teacher training programs. *Florida Journal of Educational Administration and Policy, 2*(2), 112–126.

Sireci, S., & Padilla, J. L. (2014). Validating assessments: Introduction to the Special Section. *Psicothema, 26*(1), 97–99.

Tiffin-Richards, S. P., & Köller, O. (2010). Comparison and synthesis of multiple standardsetting methods and panels. In C. Harsch, H. A. Pant & O. Köller (Hrsg.), *Calibrating standards-based assessment tasks for English as a first foreign language* (S. 107–110). Münster: Waxmann.

Wilson, M. (2005). *Constructing measures. An item response modeling approach*. New Jersey: Lawrence Erlbaum.

Zhang, S., & Stephens, V. (2013). Learning to teach English-language learners in mainstreamed secondary classrooms. *Teacher Education and Practice, 26*(1), 99–116.

Abstract: Pre-service teachers of all subject areas need learning opportunities in the field of German as a Second Language (GSL) in order to adequately facilitate second language learners in the content classroom. On the basis of a competency model including the sub-competencies *Subject-specific registers, Multilingualism* and *Didactics,* the BMBF-project *DaZKom* has conducted a validation study. This included developing a test instrument with 68 items and analyzing it with IRT-methods. The results of the validation study (N = 252) show the expected correlations between GSL competency and linguistic knowledge as well as pedagogical knowledge, yet that test measures a conceptually independent construct. Furthermore, it is found that pre-service teachers with a higher number of learning opportunities and semesters, and with German as a study subject, show a higher GSL competency.

Keywords: German as a Second Language, Teacher Education, Competency, Validation, Mathematics

Anschrift der Autor(inn)en

M.A. Svenja Hammer, Leuphana Universität Lüneburg, Institut für Bildungswissenschaft, Scharnhorststraße 1, 21335 Lüneburg, Deutschland
E-Mail: svenja.hammer@leuphana.de

M.A. Sonja A. Carlson, Universität Bielefeld, Fakultät für Erziehungswissenschaft, Universitätsstraße 25, 33615 Bielefeld, Deutschland
E-Mail: scarlson1@uni-bielefeld.de

Prof. Dr. Timo Ehmke, Leuphana Universität Lüneburg, Institut für Bildungswissenschaft, Scharnhorststraße 1, 21335 Lüneburg, Deutschland
E-Mail: tehmke@leuphana.de

Prof. Dr. Barbara Koch-Priewe, Universität Bielefeld, Fakultät für Erziehungswissenschaft, Universitätsstraße 25, 33615 Bielefeld, Deutschland
E-Mail: bkoch-priewe@uni-bielefeld.de

Dr. Anne Köker, Universität Bielefeld, Fakultät für Erziehungswissenschaft, Universitätsstraße 25, 33615 Bielefeld, Deutschland
E-Mail: anne.koeker@uni-bielefeld.de

Prof. Dr. Udo Ohm, Universität Bielefeld, Fakultät für Linguistik und Literaturwissenschaft, Universitätsstraße 25, 33615 Bielefeld, Deutschland
E-Mail: udo.ohm@uni-bielefeld.de

M.A. Sonja Rosenbrock, Universität Bielefeld, Fakultät für Linguistik und Literaturwissenschaft, Universitätsstraße 25, 33615 Bielefeld, Deutschland
E-Mail: sonja.rosenbrock@uni-bielefeld.de

M.Ed. Nina Schulze, Universität Bielefeld, Fakultät für Erziehungswissenschaft, Universitätsstraße 25, 33615 Bielefeld, Deutschland
E-Mail: n.schulze@uni-bielefeld.de

Josef Riese/Christoph Kulgemeyer/Simon Zander/Andreas Borowski/
Hans E. Fischer/Yvonne Gramzow/Peter Reinhold/Horst Schecker/
Elisabeth Tomczyszyn

Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik

Zusammenfassung: Im Hinblick auf die Verbesserung der Lehrerbildung besteht ein hohes Interesse daran, die Struktur und Genese der Kompetenzen von Lehramtsstudierenden zu beschreiben. Die Ziele des vorgestellten Projekts liegen in der Modellierung und Messung domänenspezifischer und generischer Kompetenzen, die Lehramtsstudierende der Physik im Hochschulstudium erwerben sollen. Das Modell umfasst differenzierte Facetten des physikalischen Fachwissens und des physikdidaktischen Wissens sowie das fachspezifische Erklärungswissen in unterrichtsnahen Vermittlungssituationen. Es enthält zudem Annahmen über Zusammenhänge dieser Bereiche des Professionswissens. Ausgehend vom Modell wurden schriftliche Testinstrumente und videografierte Lehr-Lern-Rollenspiele entwickelt und bei Physik-Lehramtsstudierenden eingesetzt. Die vorliegenden Ergebnisse sprechen für die Validität der genutzten Testverfahren.

Schlagworte: Professionswissen, Lehramtsstudierende, Physik, Kompetenzmessung, Validität

1. Einleitung

1.1 Problemstellung

Um die Wirksamkeit der universitären Lehrerbildung abschätzen zu können, müssen die Struktur und die Genese derjenigen Kompetenzen beschrieben werden, die Lehramtsstudierende am Ende ihres Studiums erworben haben sollten. Allerdings sind evidenzbasierte Vorschläge zur Verbesserung des Lehramtsstudiums bisher nicht ausreichend vorhanden. Large-Scale-Studien, die den Standards nationaler und internationaler Vergleichsstudien aus dem Schulbereich entsprechen, liegen in der internationalen Lehrerbildungsforschung nicht vor (vgl. Zlatkin-Troitschanskaia & Kuhn, 2010). Das gilt auch für die naturwissenschaftlichen Fächer. Zwar sind in den letzten Jahren erste Arbeiten zur Kompetenzmodellierung und -messung in der Physik entstanden, die ausgehend von der von Baumert und Kunter (2006) vorgeschlagenen Strukturierung *Fachwissen – fachdidaktisches Wissen – allgemeines pädagogisches Wissen* Kompetenzmessungen bei Lehrkräften oder Lehramtsstudierenden der Physik vorgenommen haben (vgl. z. B. die Arbeiten von Tepner et al., 2012; Riese & Reinhold, 2012; Kröger, Neumann & Petersen, 2013). Die bisherigen Arbeiten können allerdings keine über grundlegende Beschreibungen hinausgehenden Erkenntnisse liefern, da in diesen Untersuchungen hauptsächlich Gesamtscores einzelner Wissensbereiche abgebildet werden,

wie z. B. ein globaler Kennwert für *fachdidaktisches Wissen*. Dementsprechend sind bisher lediglich globale Modellprüfungen möglich, nicht jedoch differenzierte Aussagen zu Teilbereichen einzelner Dimensionen des Professionswissens. Für eine Analyse des Professionswissens von Lehramtsstudierenden ist aber eine differenziertere Modellierung wünschenswert, die z. B. die innere Struktur von *Fachwissen* und *fachdidaktischem Wissen* und die Beziehung zwischen diesen Wissensbereichen abbilden kann, um damit z. B. die Wirkung unterschiedlicher Lerngelegenheiten und längsschnittliche Entwicklungsverläufe untersuchen zu können.

1.2 Ziele der Untersuchung

Vor dem Hintergrund des oben skizzierten Problemfelds leistet das im vorliegenden Beitrag beschriebene Projekt ProfiLe-P (*„Professionswissen in der Lehramtsausbildung Physik"*, gefördert vom BMBF – FKZ 01PK11001) einen empirisch fundierten Beitrag zur Diskussion über die Wirkungen universitärer Lehrerbildung. Die Ziele von ProfiLe-P liegen in der Modellierung domänenspezifischer und generischer Wissensbereiche, die Lehramtsstudierende der Physik im Hochschulstudium als Teil ihrer professionellen Kompetenz erwerben sollten. Darüber hinaus werden die theoretischen Modelle in Messmodelle und Testinstrumente überführt, die nach mehreren Schritten der Validierung zentrale Wissensbereiche angehender Physiklehrkräfte differenzierter als bisher überprüfbar machen. Als innovatives Testformat werden praxisnahe Rollenspiele genutzt, um spezifische Anforderungen realen Lehrerhandelns in Laborsituationen abzubilden und damit, unabhängig vom empirisch schwer zugänglichen Untersuchungsfeld Schule, messbar zu machen. Mit den entwickelten Testinstrumenten wird ein Datenpool aufgebaut, der Zusammenhangsanalysen zwischen Wissensbereichen des Professionswissens sowie differenzielle Analysen ausgewählter Kompetenzen nach Lehrämtern und Studiengangsstrukturen erlaubt und darüber hinaus auf der Basis von Quasi-Längsschnitt- und echten Längsschnitterhebungen Aufschlüsse über Kompetenzentwicklungsverläufe in ausgewählten Studienabschnitten liefert.

In diesem Beitrag wird über die Kompetenzmodellierung (Projektziel 1, siehe Abschnitt 3), die Testentwicklung und die Methodik (Projektziel 2, siehe Abschnitt 4) und die Validität berichtet (Projektziel 3, siehe Abschnitt 5). Die Erhebungen zum Aufbau des gemeinsamen Datenpools (Projektziel 4) sind zurzeit (Stand August 2014) noch nicht abgeschlossen, sodass Zusammenhangsanalysen zwischen den Wissensbereichen noch ausstehen. Im Folgenden wird zunächst der theoretische Hintergrund der Kompetenzmodellierung beschrieben.

2. Theoretischer Hintergrund

2.1 Das professionelle Wissen von Lehrkräften

Der Begriff *Professionswissen* bezeichnet in der Regel spezifisches Wissen, über das die Angehörigen einer bestimmten Profession verfügen sollten (vgl. Stichweh, 1996). In Bezug auf Lehrkräfte ist dabei jenes Wissen gemeint, das den Lehrberuf auszeichnet und das als notwendig angesehen wird, die Anforderungen dieses Berufs zu bewältigen. Dem Modell von Baumert und Kunter (2006) folgend lässt es sich auf erster Strukturierungsebene in fachliches Wissen, fachdidaktisches Wissen und nicht fachbezogenes pädagogisches Wissen unterteilen.

Fachliches Wissen (FW) bezieht sich hierbei sowohl auf Wissen über die Inhalte und Methoden des jeweiligen Fachs als auch auf epistemologisches Wissen über das Fach, in dem eine Lehrperson unterrichtet. Die historisch generierte inhaltliche und logische Struktur der Fachdisziplin (insbesondere die üblicherweise unterschiedenen Inhaltsgebiete) wird dabei in zentralen Lehrwerken und in Curricula von Studienprogrammen abgebildet. Obwohl FW sowohl eine Grundvoraussetzung für erfolgreichen Fachunterricht darstellt als auch den Erwerb von fachdidaktischem Wissen beeinflusst (vgl. z. B. Baumert & Kunter, 2006), fand es in der Untersuchung des Professionswissens von Lehramtsstudierenden oder Lehrkräften bisher wenig Berücksichtigung (vgl. Fischer, Borowski & Tepner, 2012).

Fachdidaktisches Wissen (FDW) wiederum bezeichnet einen für den Lehrberuf spezifischen Wissensbereich, der häufig in Anlehnung an das Wissenskonzept des *pedagogical content knowledge* (PCK) von Shulman (1986) definiert wird. Danach benötigen Lehrkräfte spezifisches Wissen, das sie dazu befähigt, fachliche Inhalte adressatenbezogen zu strukturieren und darzustellen. In den letzten Jahren wurden zahlreiche Modellierungen vorgenommen, in denen das ursprüngliche Konstrukt unterschiedlich ausdifferenziert wurde, allerdings existiert für FDW in den Naturwissenschaften bislang weder national noch international eine einheitliche, umfassende und auf ein Rahmenmodell zum Professionswissen abgestimmte Modellierung (vgl. Abell, 2007; Park & Chen, 2012; Fischer et al., 2012). Die meisten Untersuchungen betrachten jedoch Wissen über Schülerkonzeptionen bzw. -vorstellungen und Wissen über Lehrstrategien und Darstellungsformen als zentrale Elemente des FDW (vgl. Gramzow, Riese & Reinhold, 2013). Kulgemeyer und Schecker (2013) sprechen darüber hinaus von Erklärungswissen (EW), das Lehrpersonen für eine sachgerechte und schülergemäße Kommunikation naturwissenschaftlicher Sachverhalte benötigen. EW weist demnach starke Bezüge zum FDW und zum FW auf, wird aber von den Autoren als eigenständiger Wissensbereich betrachtet, da es sich ausschließlich auf Anforderungen in einer konkreten, unterrichtsbezogenen Handlungssituation bezieht.

Pädagogisches Wissen (PW) bezeichnet schließlich fachunabhängiges Wissen über pädagogische und allgemeindidaktische Konzepte und Inhalte. Da wir in der im Beitrag vorgestellten Untersuchung auf physikbezogenes Professionswissen fokussieren, wird PW im Folgenden nicht weiter ausgeführt.

2.2 Untersuchungen zum fachbezogenen Professionswissen von Lehrkräften

In der nationalen Literatur herrscht weitgehend Konsens im Hinblick auf die oben beschriebene Grobstruktur des Professionswissens von (angehenden) Lehrkräften. Dennoch wurden bisher häufig nur einzelne Bereiche unabhängig voneinander untersucht (Fischer et al., 2012). Auch für einzelne Fächer liegen kaum Erkenntnisse zur Struktur des fachbezogenen Professionswissens vor, die mehr als globale Zusammenhänge (wie z. B. bei Riese & Reinhold, 2012) zwischen den drei Bereichen aufklären. Allerdings ist zu vermuten, dass einige Aspekte des FDW (z. B. Wissen über typische fachliche Schüler(fehl)konzepte) ein tieferes Verständnis physikalischer Aspekte erfordern als andere (z. B. Wissen über fachdidaktische Unterrichtsmodelle), sodass globale Korrelationen die Zusammenhänge nur oberflächlich aufklären. Mangels empirisch bislang nicht trennbarer Unterfacetten des FDW können vorliegende Studien zum Zusammenhang auf Teilskalenebene jedoch keine Aussage machen (vgl. Gramzow et al., 2013). Weiter ist festzustellen, dass das Professionswissen von (angehenden) Lehrkräften bislang fast ausschließlich mit schriftlichen Leistungstests erfasst wurde, wobei unklar ist, inwieweit die erfassten Wissenselemente die Qualität der Handlung von Lehrpersonen valide prognostizieren können (Vogelsang & Reinhold, 2013).

2.3 Konzeptionen des Professionswissens in den Naturwissenschaften

Im Vergleich zu Modellen naturwissenschaftlicher Kompetenz auf Schülerebene (wie z. B. bei Pant et al., 2013) müssen Modelle des Professionswissens von Lehrkräften deutlich umfangreicher angelegt sein, da mehr Wissensbereiche einzubeziehen sind. Im Folgenden werden typische Modellierungsansätze zum FW und FDW skizziert, die in bisherigen Studien als Ausgangsbasis für Itementwicklungen genutzt wurden.

Modelle physikalischen Fachwissens

Empirische Belege für eine bestimmte inhaltliche Ausprägung oder Dimensionierung physikalischen Fachwissens oder für die Anzahl der Facetten in einzelnen Dimensionen liegen bisher nicht vor. Häufig finden sich jedoch mehrdimensionale konzeptionelle Modelle. Üblicherweise beschreiben Autoren dabei zum einen eine inhaltliche Dimension, welche typische Themen der Physik unterscheidet (z. B. Kröger et al., 2013; Riese & Reinhold, 2012). Zum anderen werden Fachstufen unterschieden, die sich auf typische Lerngelegenheiten des Wissenserwerbs beziehen und die in den verschiedenen Untersuchungen teilweise Schwierigkeit generiert haben. Im Bereich Primarstufe unterscheidet Ohle (2010) zur Konstruktion eines Testinstruments für das Fachwissen z. B. die Fachstufen *Grundschule, Mittelstufe* sowie *Universität.* Daneben findet sich häufig eine dritte Teildimension, die es ermöglichen soll, Testitems a priori nach ihrer Aufgabenschwierigkeit anzuordnen und Unterschiede im Wissen stärker auf die Struktur des Wissens selbst zu beziehen. Beispielsweise unterscheiden Tepner et al. (2012) zwischen *Wissensarten* (deklaratives, prozedurales und konditionales Wissen) und Woitkowski,

Riese & Reinhold (2011) zwischen unterschiedlichen *Komplexitäten* (Fakten, Prozessbeschreibungen, lineare Kausalität, multivariate Interdependenz).

Modelle fachdidaktischen Wissens

Ebenso wie beim FW liegen in der Literatur zum FDW konzeptionelle Modelle unterschiedlicher Dimensionierung vor. Alle Modelle unterscheiden explizit fachdidaktische Inhalte, welche bezeichnet werden: als *Inhalte* (z. B. Schülerkognition, Instruktionsstrategien, Curriculum, Assessment; bei Kröger et al., 2013), als *Facetten* (Schülervorstellungen, Modelle/Konzepte, Experimente; bei Tepner et al., 2012) oder als *Kategorien* (Wissen über Schülervorstellungen, über das Curriculum und über Schwierigkeiten; vgl. Olszewski, 2010). Darüber hinaus unterscheiden einige Studien analog zu den entsprechenden Modellen des FW auch für FDW Wissensarten und Fachinhalte, um eine strukturelle Einheitlichkeit zu betonen (z. B. Tepner et al., 2012 oder Kröger et al., 2013; siehe oben).

3. Modellierung des Professionswissens von Lehramtsstudierenden der Physik

Das im Forschungsverbund ProfiLe-P entwickelte Rahmenmodell des Professionswissens angehender Physiklehrkräfte unterscheidet auf erster Strukturierungsebene zwischen physikalischem Fachwissen (FW), eher deklarativen und analytischen Aspekten fachdidaktischen Wissens (FDW) und eher prozeduralen Wissensaspekten, welche beim Erklären physikalischer Sachverhalte im Physikunterricht zum Tragen kommen. Von diesem fachspezifischen Erklärungswissen (EW) wird hypothetisch erwartet, dass es Auswirkungen auf das Erklärungshandeln in realen Vermittlungssituationen hat. Der Fokus wurde auf das Erklären physikalischer Sachverhalte gelegt, da es sich beim Erklären um eine wichtige Standardoperation im naturwissenschaftlichen Unterricht handelt (Geelan, 2012). Zudem stellt es Bezüge zwischen deklarativen und prozeduralen Wissensaspekten her. So kann beispielsweise untersucht werden, inwiefern deklarative Wissensaspekte, z. B. Wissen über typische Verständnisprobleme bei Lernenden, in einer unterrichtlichen Handlungssituation relevant werden (vgl. auch Vogelsang & Reinhold, 2013). Abbildung 1 setzt die im Vorhaben untersuchten Bereiche des Professionswissens (mit ihren jeweiligen Facetten) in Beziehung.

In allen drei Wissensbereichen konzentrieren wir uns entsprechend der Leitlinie des BMBF-Rahmenprogramms „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“ (vgl. Blömeke & Zlatkin-Troitschanskaia, 2013) auf universitär erwerbbares physikbezogenes Professionswissen. Als normativ-präskriptive Orientierungspunkte bei der Entwicklung der jeweiligen Kompetenzmodelle wurden u. a. die KMK-Standards für die Lehrerbildung (KMK, 2008; Fachprofil Physik), das Kerncurriculum der Gesellschaft für Fachdidaktik (GFD, 2004), der Fachdidaktikteil der Studie der Deutschen Physikalischen Gesellschaft zur Ausbildung für das Lehramt Physik (DPG, 2014), aber auch die Fachphysik- und Physikdidaktikcurricula der Universitä-

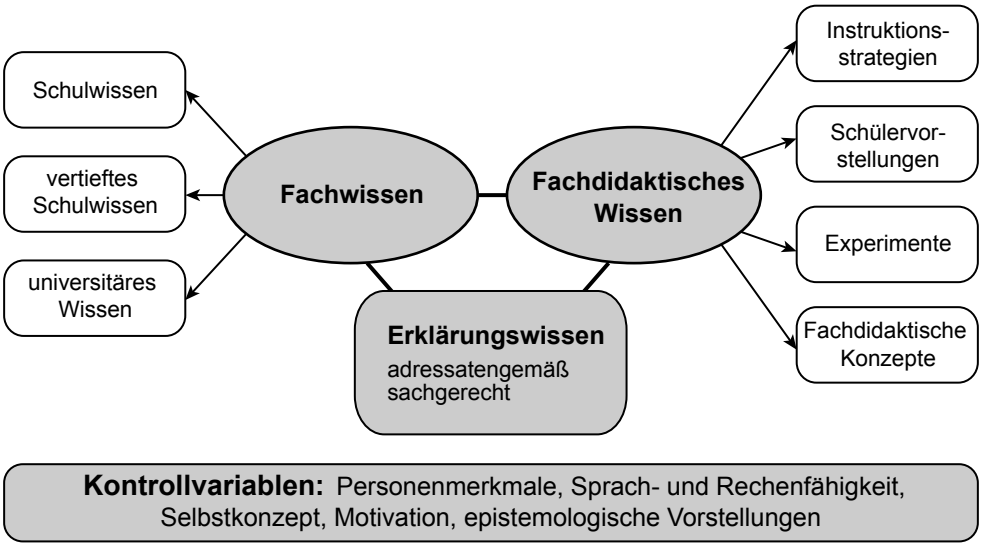


Abb. 1: Mögliche Zusammenhänge zwischen unterschiedlichen Bereichen des Professionswissens

ten des Verbunds sowie Erkenntnisse zum Eingangswissen von Studierenden (Riese & Reinhold, 2012) herangezogen.

Inhaltlich beziehen sich alle Modellierungen auf den Bereich Mechanik, da die Erhebungen mit dem ersten Studiensemester beginnen, in dem im Wesentlichen Mechanik gelehrt wird. Zudem sind Schülervorstellungen und Schülerlernprozesse in der Mechanik gut erforscht (z. B. Schecker, 1985) und daher für die Entwicklung von Items zum FDW und zum EW von Studierenden gut geeignet. Schließlich korreliert FDW in Mechanik nach Borowski & Riese (2010) hoch mit FDW in Elektrizitätslehre.

Im Folgenden werden die Binnenstrukturen der drei Modelle vorgestellt.

3.1 Fachwissen

Das Modell für das Fachwissen unterscheidet die Fachstufen *Schulwissen, vertieftes Schulwissen* und *universitäres Wissen* (Abb. 1, vgl. auch z. B. COACTIV-Studie). *Schulwissen* umfasst Inhalte der Richtlinien und Lehrpläne der an diesem Projekt teilnehmenden Bundesländer, mit Beschränkung auf den Grundkurs der Oberstufen. Das *vertiefte Schulwissen* modelliert folgende Fähigkeiten: (1) verschiedene Wege zur Lösung einer Aufgabe identifizieren und anwenden, (2) Lösung einer Aufgabe aus theoretischer Sicht planen, (3) Randbedingungen einer Schulaufgabe erkennen, (4) Aufgaben zielgruppengerecht vereinfachen, (5) Zusammenhänge, Gemeinsamkeiten und Unterschiede physikalischer Phänomene erkennen. *Universitäres Wissen* enthält Inhalte, die an Universitäten für Abschlussprüfungen im Fach vorausgesetzt werden. Bei diesen drei

Fachstufen handelt es sich um qualitativ unterschiedliche Bereiche, von denen erwartet wird, dass sie empirisch zu trennen sind und über mittlere Korrelationen zusammenhängen. Um zu gewährleisten, dass Aufgaben unterschiedlicher Schwierigkeit erstellt werden, wurde die Komplexität (Schoppmeier, Borowski & Fischer, 2013) als weitere Teildimension bei der Entwicklung berücksichtigt, wobei zwischen *Fakten benennen, Zusammenhänge zwischen mehreren Fakten herstellen* und dem *Umgang mit übergeordneten physikalischen Konzepten* unterschieden wird. Bei Aufgaben zu übergeordneten physikalischen Konzepten muss ein grundlegendes physikalisches Prinzip (z. B. Basiskonzept Energie) verstanden, in einer gegebenen Situation erkannt oder auf neue Situationen angewandt werden.

3.2 Fachdidaktisches Wissen

Ausgehend von einem fachdidaktischen Modell, das unter Nutzung allgemeiner Standards und normativer Leitbilder der Physiklehrerbildung (z. B. DPG, 2014; Korneck, Lamprecht, Wodzinski & Schecker, 2010) entwickelt wurde und den gesamten Bereich des FDW umspannt (detailliert in Gramzow et al., 2013), wurden vier Facetten für ein separates Modell zur Itementwicklung ausgewählt, um den Testumfang auf ein praktikables Maß zu begrenzen. Hierbei handelt es sich um die fachdidaktischen *Facetten Instruktionsstrategien, Schülervorstellungen, Experimente und Vermittlung eines angemessenen Wissenschaftsverständnisses* sowie *Fachdidaktische Konzepte* (vgl. Abb. 1), die im Sinne einer empirisch zu prüfenden Binnenstruktur des FDW zu verstehen sind. Es wird erwartet, dass diese Facetten für das Erklärungswissen (EW) besonders relevant sind (siehe auch 3.3). Als weitere Teildimension wurden *Kognitive Anforderungen* in den Stufungen „Reproduzieren“, „Analysieren“ und „Anwenden“ modelliert (vgl. Anderson & Krathwohl, 2001), um das Spektrum der Anforderungen breit erfassen zu können.

3.3 Erklärungswissen

Das Modell des Erklärungswissens bezieht sich auf unterrichtsnahe, dialogische Erklärungssituationen von Physik (in Anlehnung an Kulgemeyer & Schecker, 2013). Es berücksichtigt als grundlegende Anforderungen die Adressatengerechtheit und die Sachgemäßheit einer Erklärung. Dabei werden folgende Variablen unterschieden: Mathematisierungsgrad, gewählte Beispiele bzw. Analogien, Sprachniveau und Darstellungsformen. Diese Variablen sind nicht im Sinne einer empirisch zu prüfenden Binnenstruktur des EW zu verstehen, sondern als Gestaltungsmittel von Erklärungen, die für die Analyse der Güte von Erklärungen verwendet werden.

In dem grundlegenden Modell wird davon ausgegangen, dass eine Erklärung ein sprachlicher Akt ist und *Erklären* den Prozess der adressatengemäßen und sachgerechten Erstellung und Modifikation von *Erklärungen* bedeutet. Der Prozess läuft im ein-

fachsten Falle wie folgt ab: Es wird eine initiale erklärende Darstellung des Sachverhalts gewählt, die nach Verständnisrückmeldung (verbal oder nonverbal) durch den Adressaten entsprechend modifiziert werden muss. Dabei sind es gerade die in 3.2 genannten vier Facetten des FDW, die eine adressatengemäße und sachgerechte Erstellung bzw. Modifikation der Variablen ermöglichen. So kann beispielsweise das Sprachniveau von der Fach- zur Alltagssprache wechseln oder die Darstellungsform von einem logischen Diagramm zu einem realen Foto. Analogien bzw. Beispiele können sich auf unterschiedliche Gegenstandsbereiche beziehen, z. B. auf fachnahe oder alltagsnahe. Der Mathematisierungsgrad kann wechseln, indem physikalische Größengleichungen expliziert oder verbal umschrieben werden. Gütekriterien für Erklärungen werden z. B. von Brown (2006) und Wellenreuther (2005) beschrieben, etwa „Evaluation des Verständnisses“. Das Modell ist die Adaption eines Modells, das empirisch erfolgreich verwendet wird, um physikbezogene Kommunikation auf Schülerebene zu beschreiben (detailliert in Kulgemeyer & Schecker, 2013).

4. Testentwicklung und Methodik

4.1 Messinstrumente

Auf der Basis der drei oben skizzierten Teilmodelle wurden sowohl schriftliche Leistungstests als auch ein qualitatives, videobasiertes Instrument entwickelt, mit dem die Ausprägung des Erklärungswissens (EW) gemessen wird. Die Inhalte der Tests zur Erhebung von FW und FDW sind dabei auf die Thematiken des Tests zum EW abgestimmt. Alle Tests adressieren den Fachinhalt Mechanik (vgl. Abschnitt 3). Dabei stand bei allen Tests das Ziel im Vordergrund, Strukturen der einzelnen Konstrukte und Zusammenhänge zu Bedingungsfaktoren (z. B. Lerngelegenheiten) untersuchen zu können, nicht aber die Ermöglichung von Individualdiagnostik.

Fachwissen

Für den FW-Test wurden, ausgehend vom in 3.1 beschriebenen Modell, insgesamt 143 geschlossene Aufgaben im Multiple-Choice-Format (Single Select) entwickelt (Bsp. siehe Abb. 2), sodass für die finale Testzusammenstellung noch ausreichend Kürzungsspielraum vorhanden war. Dabei wurden Items zu allen Subdimensionen des Modells erstellt, um das Konstrukt bestmöglich abzudecken.

Der Test wurde in einer ersten Feldstudie im Multimatrixdesign mit 395 Studierenden des Lehramts Physik (N = 120), des Physik-Monobachelors (N = 213; Vergleichsgruppe) und sonstigen Physikstudierenden (N = 62; z. B. Mathematikstudierende mit Physik als Nebenfach) eingesetzt (EAP/PV-Reliabilität .76; Varianz 0.93; bei 127 der 143 Aufgaben gilt $0.8 <$ gewichteter MNSQ < 1.2 und $-1.9 < T < 1.9$). Auf dieser Datengrundlage wurden 40 Aufgaben für den weiteren Einsatz ausgewählt, die in der Pilotierung möglichst gute Kennwerte aufwiesen als auch einen möglichst großen Ausschnitt der Fähigkeitsskala abdeckten, um zu gewährleisten, dass der finale Test während des

Betrachten Sie ein Pendel (z. B. ein Federpendel). Damit dieses Pendel harmonisch schwingt, muss die zurücktreibende Kraft einer ganz bestimmten Gesetzmäßigkeit gehorchen.
Im Folgenden bezeichnen x die Auslenkung des Pendels, $\dot{x}$ die Geschwindigkeit des Pendels, k eine Konstante mit $k < 0$ und F die zurücktreibende Kraft.
Welche Aussage ist richtig?

Es muss gelten:

1) $F = k$
2) $F = k \cdot x$
3) $F = \frac{1}{2} k \cdot x^2$
4) $F = k \cdot \dot{x}$

Abb. 2: Beispielitem FW zum Schulwissen, einen Fakt benennen (Lösung: 2)

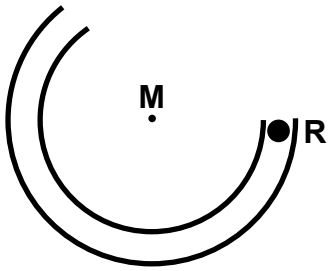
kompletten Studiums eingesetzt werden kann. Weitere Kriterien für die Auswahl waren, dass zentrale für die Mechanik relevante Themenbereiche (z. B. Energie/Arbeit/Leistung) erfasst werden und die Inhalte das notwendige FW zur Beantwortung der fachdidaktischen Aufgaben sowie der Erklärungssituationen abdecken. Zusätzlich wurde berücksichtigt, dass die drei Fachstufen der theoretisch angenommenen inneren Struktur des Fachwissens (vgl. Abschnitt 3.1) getestet werden können (Schulwissen: N = 17 Items; universitäres Wissen: N = 12 Items; vertieftes Schulwissen: N = 11 Items).

Insgesamt haben die Studierenden für die Bearbeitung des finalen Tests 60 Minuten Zeit, wobei der Rechenaufwand zur Auswahl der richtigen Lösungen ohne Taschenrechner zu leisten ist. Zur Erfassung des Einflusses der mathematischen Fähigkeiten auf das Lösen der Physikaufgaben wurde zudem ein Rechentest entwickelt, der die mathematischen Anforderungen der Physikfachwissen-Items separat darstellt. Hierbei wurde auf die mathematischen Inhaltselemente Terme, Trigonometrie, Vektoren und Matrizen, Differenzieren und Integrieren fokussiert.

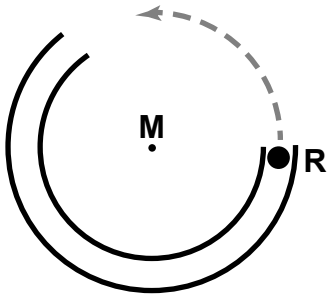
Fachdidaktisches Wissen

Im Zuge der Entwicklung des FDW-Tests wurden zunächst die Subdimensionen des Modells aus 3.2 im Sinne einer deduktiven Testentwicklung ausdifferenziert, um Kerninhalte für einzelne Items zu generieren, sodass eine bestmögliche Abbildung des Konstrukts gewährleistet wird. Davon ausgehend wurden zielgerichtet Items konstruiert, die bestimmte Kerninhalte adressieren sollen. In diesem Zusammenhang konnten einige Items, sofern sie eine Passung zum Modell aufwiesen, von Riese und Reinhold (2012) adaptiert werden. Alle Items wurden anschließend im gesamten Projektteam intensiv diskutiert und überarbeitet. Insgesamt wurden so 50 Aufgaben mit 91 sowohl offenen als auch geschlossenen (Multiple-Choice-Format, Multiple Select) Items konstruiert (Schülervorstellungen: 16 Aufgaben; Instruktionsstrategien: 7 Aufgaben; Experimente

Schüler sollen folgende Situation betrachten: Ein Ball rollt in der dargestellten Rinne (Draufsicht) und verlässt diese am Punkt R.



a) Schüler A zeichnet folgende Bahn, die der Ball nach Verlassen der Rinne beschreiben soll:



Lösung von Schüler A

Angenommen, der Schüler versteht die Zeichnung korrekt als Draufsicht: Welche fachlich nicht korrekte Vorstellung des Schülers A liegt bei der gezeichneten Bahnkurve zugrunde?

Abb. 3: Beispielitem FDW zu Schülervorstellungen zur Kreisbewegung

und Vermittlung eines angemessenen Wissenschaftsverständnisses: 12 Aufgaben; Fachdidaktische Konzepte: 15 Aufgaben). Diese wurden auf zwei Testhefte für jeweils 90 Minuten Bearbeitungszeit (inklusive Angaben zur Demografie) aufgeteilt, wobei 14 Aufgaben in beiden Testheften enthalten sind (Multi-Matrix-Design).

Abbildung 3 zeigt ein Item, welches sich auf die kognitive Anforderung „Analysieren“ und die Facette „Schülervorstellungen“ bezieht. Eine (falsche) Schülerantwort muss dahingehend analysiert werden, welche Schüler(fehl)vorstellung der Antwort zugrunde liegt. Die Idee zu diesem Item lieferte der Force Concept Inventory Test (vgl. Hestenes, Wells & Swackhamer, 1992).

Da ein Testheft mit einer Bearbeitungsdauer von 90 Minuten für den Einsatz im Verbund mit den anderen Tests zu umfangreich war, blieb Spielraum, Items mit der besten Inhaltsvalidität auszuwählen (entsprechende Analysen siehe 5.2). Für die Hauptstudie wurden schließlich 26 Aufgaben (43 Items) übernommen, die in einem Testheft mit einer Bearbeitungszeit von 75 Minuten (inklusive Demografie) zusammengefasst wurden.

Erklärungswissen

Das EW wird in einer standardisierten, interaktiven Testsituation erhoben, bei der die Probanden ein vorgegebenes physikalisches Phänomen einer Schülerin bzw. einem Schüler aus der zehnten Klasse erklären sollen, z. B. das Gefühl der Schwerelosigkeit bei der Achterbahnfahrt (vgl. Kulgemeyer & Schecker, 2013). In ihrer zehnminütigen Vorbereitungszeit werden den Erklärenden Anschauungsmaterialien zur Verfügung gestellt, die sie in der Erklärung nutzen können, beispielsweise Diagramme, Bilder oder einschlägige Formeln (siehe Abb. 4).

Die Erklärungen (10 Min. Dauer) werden dann videografiert. Den Probandinnen und Probanden ist nicht bekannt, dass die Schülerinnen und Schüler vorher darauf trainiert worden sind, sich in allen Erklärungssituationen in einer bestimmten Weise zu verhalten, insbesondere gleichartige Fragen zu stellen, z. B. „Gibt es dafür ein Beispiel?“.

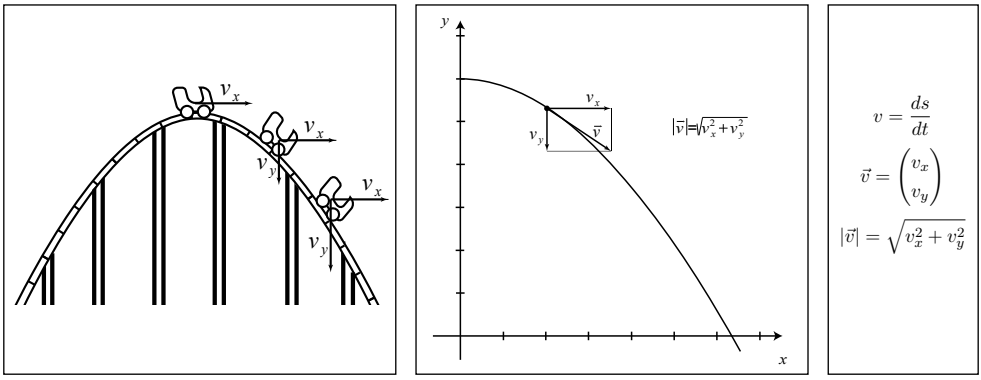


Abb. 4: Beispiele für zur Verfügung stehende Darstellungsformen zum Thema „Schwerelosigkeit in der Achterbahn“

Durch solche Prompts werden in allen Erklärungssituationen vergleichbare Anforderungen an die Variation der Variablen Mathematisierungsgrad, Beispiele, Sprachniveau und Darstellungsformen aus dem oben genannten Modell gestellt. Der Einsatz dieser Prompts wird mit den Schülerinnen oder Schülern in nachgestellten Erklärungsszenen geübt und anhand von Videoaufzeichnungen besprochen, bis das angestrebte Verhalten zur Routine geworden ist.

Zur weiteren Auswertung wurde ein Kategoriensystem mit 45 mittel- bis niedrig-inferent einzuschätzenden Kategorien mit Ankerbeispielen entwickelt, z. B. „Erklärer gibt ein Zahlenbeispiel". Diese Kategorien wurden mit qualitativer Inhaltsanalyse (Mayring, 2003) gewonnen. Ziel war dabei das ökonomische Kodieren am Videomaterial ohne vorherige Transkribierung. Das Verfahren war dabei wie folgt: Ein Startsatz an Überkategorien ergab sich deduktiv aus dem Modell für EW, indem Variationen in den im Modell angenommenen Variablen kategorisiert wurden (z. B. Variationen im Sprachniveau, siehe 3.3). Diese Überkategorien wurden am Material induktiv in Unterkategorien verfeinert, wozu Ankerbeispiele festgehalten und daraus Generalisierungen erstellt wurden, die zur Unterkategorie weiterentwickelt wurden. So gab es unterschiedliche Arten, das Sprachniveau zu variieren, z. B. indem auf lexikalischer Ebene Fachbegriffe direkt nach Nennung erläutert oder ganz vermieden wurden und eine Erläuterung auf alltagssprachlichem Niveau erfolgte. Zudem wurde das Auftreten von aus der Literatur bekannten Qualitätskriterien für Erklärungen deduktiv mitkodiert (z. B. Evaluation des Verständnisses, siehe 3.3). Auch diese Kategorien dienten wieder als Oberkategorien, die in Unterkategorien differenziert wurden. Die Evaluation des Verständnisses konnte z. B. erfolgen, indem direkte Fragen oder Aufgaben gestellt wurden, die den erklärten Inhalt thematisierten. Es zeigte sich bei der Analyse der Videos durch einen Rater, dass die Kategorien und damit das Modell erschöpfend alle Variationen beschreiben, die die Probanden in ihren Erklärungen vornehmen, um deren Verständlichkeit zu verbessern. Alle Veränderungen konnten durch den Rater Kategorien zugeordnet werden. Zur Überprüfung der Reliabilität der Einstufungen wurden Interraterreliabilitäten bei den Kategorien berechnet. Dabei werden bei zwei Ratern bislang Übereinstimmungen zwischen 73 % und 97 % erreicht. Durch anschließende Kommunikation konnte in allen Fällen Einigkeit erzielt werden, weitere Berechnungen laufen.

In einer ersten Vorstudie wurden fünf physikalische Themen zur Erklärung eingesetzt. Hiervon wurden drei im Anschluss weiter verwendet. Kriterien für die Auswahl waren neben curricularer Validität (siehe 5.3) vor allem ein gemessen am fachlich-physikalischen Inhalt vergleichbarer Schwierigkeitsgrad. Dazu wurden die Videos gesichtet und Themen, bei denen gehäuft Probleme auftraten, nach einvernehmlicher Diskussion von drei Personen im Projektteam gestrichen.

Miterhoben werden zudem Erklärungserfahrung, erklärungsbezogene Selbstwirksamkeitserwartung und Interesse am Erklären. Dazu wurden eigene Fragebögen mit Likert-Skalen entwickelt, für die Selbstwirksamkeitserwartung ergibt sich ebenso eine reliable Skala (Cronbachs $\alpha = 0.71$) wie für das Interesse am Erklären (Cronbachs $\alpha = 0.79$).

4.2 Datenerhebungen

Die beschriebenen Testinstrumente wurden für den Einsatz bei Studierenden der Lehrämter an Gymnasien/Gesamtschulen und Haupt-/Real-/Gesamtschulen entwickelt. In einer Begleiterhebung zum FW, FDW und EW werden neben allgemeinen Personenmerkmalen (PM) die Rechenfähigkeit, kognitiv-sprachliche Fähigkeiten, der Studienfortschritt (SWS und Credit Points) als Maß für bisherige Lerngelegenheiten sowie fachbezogenes Interesse, fachbezogenes Selbstkonzept, die Fähigkeit zur Perspektivenübernahme und epistemologische Überzeugungen erhoben. Die Daten werden für die Aufklärung bestimmter Fähigkeitsausprägungen in den Bereichen des FW, FDW und EW genutzt.

In einer ersten Erhebungswelle, die als Querschnittserhebung über alle Studiensemester primär zur Generierung von Validitätsargumenten durchgeführt wurde, wurden die jeweiligen (noch zu umfangreichen) Tests zunächst in Stichproben eingesetzt, die nur zum Teil zusammenhängen, um die Testzeit der Probanden im erträglichen Rahmen zu halten. Die Probandenzahlen der einzelnen Teiluntersuchungen sind in den entsprechenden Teilen von Abschnitt 5 aufgeführt.

In der laufenden Haupterhebung, bei der alle Instrumente in einer gemeinsamen Stichprobe eingesetzt werden, werden ca. 500 Testpersonen unterschiedlicher Studiensemester an den vier Universitäten des Forschungsverbunds jeweils schriftlich zum FW und FDW getestet, davon sollen ca. 200 in etwa gleich verteilt über die verschiedenen Studiensemester zusätzlich zum EW getestet werden. Der Datenpool ermöglicht somit bereichsübergreifende Zusammenhangsanalysen. Die Datenerhebung erfolgt dabei mittels einer Kombination aus Quasi-Längsschnitt und echtem Längsschnitt (Erhebungen vom 1. bis 3. Semester, Erhebungen vom 3. bis 5. Semester) sowie eines Querschnitts bei fortgeschrittenen Studierenden.

4.3 Testtheorie

Um einen Zusammenhang zwischen den Rohdaten und den daraus resultierenden Leistungsscores herzustellen, werden in ProfiLe-P verschiedene testtheoretische Ansätze verwendet. Für die Tests zur Erfassung des FW und des FDW werden Verfahren der probabilistischen Testtheorie genutzt, da diese gegenüber der klassischen Testtheorie mehr Möglichkeiten bieten (z. B. Stichprobenunabhängigkeit, Ermöglichung einer großen Itemzahl zur angemessenen Abdeckung der Kompetenzmodelle im Multi-Matrix-Design; vgl. Rost, 2004; Moosbrugger & Kelava, 2012). Konkret werden im Falle von FW und FDW ein- und mehrdimensionale Rasch-Modelle verwendet, um Zusammenhänge zwischen der Lösungswahrscheinlichkeit der Testitems und den verschiedenen Parametern herzustellen. Zur Überprüfung von Hypothesen zur Dimensionalität der erfassten Konstrukte werden unterschiedliche Modelle in Bezug auf die Passung zum Datensatz mittels der Informationskoeffizienten AIC und BIC sowie mithilfe des Chi-Quadrat-Differenztests verglichen (ebd.).

Auf eine Verwendung des 2-pl-Birnbaum-Modells wird aufgrund der Stichprobengröße in den hier berichteten Studien zunächst verzichtet, da in der Literatur eine Verwendung nur dann als sinnvoll erachtet wird, wenn eine große Stichprobe zur Verfügung steht, die das gesamte Fähigkeitsspektrum hinreichend abdeckt (vgl. Moosbrugger & Kelava, 2012). Diese Bedingungen sind jedoch für die berichteten Studien zur Sammlung von Validitätsargumenten nicht gegeben, da z. B. mit dem FDW-Test nur gut 200 Probanden befragt werden konnten, wobei nicht genug Personen im oberen Fähigkeitsbereich zu verorten sind.

Im Hinblick auf die deutlich aufwendiger zu generierenden Daten des Tests zur Erfassung des EW mittels Videografie (vgl. 4.2) stehen bisher nicht genug Probanden zur Verfügung, um Methoden der probabilistischen Testtheorie nutzen zu können. Dementsprechend erfolgt für diesen Fall die Nutzung der klassischen Testtheorie, um die Fähigkeitsparameter zu generieren.

5. Nachweis von Validität der Erhebungsinstrumente

Zur Validierung verfolgt ProfiLe-P einen Argument-based-Ansatz, der sich auf die Interpretation eines Testwerts bezieht, wie ihn Jenßen, Dunekacke und Blömeke (2015) im Einleitungsbeitrag dieses Beihefts vorschlagen (vgl. auch Kane, 2013). Dabei ist zu beachten, dass die Validität einer Testwertinterpretation – z. B. bzgl. der Verallgemeinerung eines Testergebnisses über die spezifischen Testinhalte hinaus – zwar durch beigebrachte Validitätsargumente unterstützt oder zurückgewiesen, nicht jedoch abschließend belegt werden kann (vgl. ebd.). In diesem Sinne wurden mehrere Teiluntersuchungen zur Generierung von Validitätsargumenten durchgeführt, die im Folgenden erläutert werden. Dabei sollen die entwickelten Testinstrumente primär eine Testwertinterpretation legitimieren, die auf die Erklärung von Leistungen abzielt (angelehnt an die Unterscheidung von Hartig, Frey & Jude, 2012).

Die in ProfiLe-P durchgeführten Validierungsstudien werden im Folgenden anhand der methodischen Kriterien *Inhaltsvalidität, Konstruktvalidität* und *Kriteriumsvalidität* gegliedert, welche im Bereich der Leistungsmessung üblicherweise unterschieden werden (vgl. Einleitungsbeitrag von Jenßen et al., 2015). Um die Probanden nicht über Gebühr zu beanspruchen, wurden die unterschiedlichen Validierungsstudien nicht in einer gemeinsamen Stichprobe, sondern für die Teilprojekte getrennt durchgeführt.

5.1 Fachwissen

Bezogen auf die Modelldimension „Fachstufen“ (Schulwissen, vertieftes Schulwissen, universitäres Wissen) erfolgte die curriculare Validierung des FW-Tests auf der Basis von Lehrbuch-Analysen für Schulen und Universitäten, Analysen der Vorlesungsinhalte für die Einführungsvorlesungen und unter Berücksichtigung der Richtlinien und Lehrpläne an weiterführenden Schulen. Es wurden dazu die Anforderungen und Inhalte aller

Einführungsvorlesungen der beteiligten Universitäten analysiert, die Unterschiede der Konzepte (z. B. die Energieerhaltung) auf den verschiedenen Fachstufen herausgearbeitet und die Aufgaben mit Fachphysikern und Fachdidaktikern diskutiert. Zudem wurde zur Inhaltsvalidierung mithilfe einer Konsenskodierung durch Lehramtsstudierende der Physik höherer Semester noch einmal geprüft, ob die Aufgaben als prototypisch zu den intendierten Teildimensionen des Modells konstruiert wurden. Hierbei wurde festgestellt, dass für die Fachstufe *vertieftes Schulwissen* von den in 3.1 benannten Fähigkeiten nur Aufgaben zu den Fähigkeiten 3 (Randbedingungen einer Schulaufgabe erkennen) und 5 (Zusammenhänge, Gemeinsamkeiten und Unterschiede physikalischer Phänomene erkennen) konstruiert wurden.

Um Hinweise zur Konstruktvalidität zu erhalten, wurde erstens die Dimensionalität der hypothetisch angenommenen inneren Struktur des Modells zum Fachwissen (wie in 3.1 beschrieben) geprüft und zweitens untersucht, in welcher Form theoretisch angenommene Zusammenhänge mit anderen Variablen vorliegen (Verortung des Konstrukts in einem nomologischen Netzwerk mittels Korrelationsanalysen; vgl. auch Einleitungsbeitrag von Jenßen et al., 2015). Dazu wurden Erhebungen an bislang vier Standorten mit insgesamt 469 Studierenden durchgeführt, davon 186 Lehramtsstudierende, 222 Physik-Monobachelorstudierende und 61 sonstige Studierende der ersten vier Semester.

Zunächst zeigt sich hier beim Vergleich des 1D-Modells der Raschanalyse (EAP/PV-Reliabilität: 0.83; Varianz: 0.90) mit dem 3D-Modell, dass das 3D-Modell zu bevorzugen ist. Die EAP/PV-Reliabilitäten für die drei Fachstufen sind gut, die Varianzen akzeptabel (vgl. Tab. 1), und die Unterschiede zwischen den Modellen sind signifikant (siehe Abb. 5).

Weiter wurde untersucht, welcher Zusammenhang zwischen der Personenfähigkeit im FW-Test mit Rechenfähigkeiten und Zeugnisnoten besteht. Dabei zeigte sich eine hohe Korrelation zwischen FW-Test und Rechentest (siehe Tab. 2), wobei die Korrelation vom Schulwissen über das vertiefte Schulwissen zum universitären Wissen hin abnimmt. Einschränkend muss angemerkt werden, dass die Korrelationen nur für die Hälfte der Stichprobe berechnet werden konnten, da die Codes der Studierenden, die sowohl am Fachwissens- als auch am Rechentest (zusammen mit dem FDW-Test zu einem anderen Zeitpunkt) teilgenommen haben, oft nicht übereinstimmten.

Bei den Noten zeigt sich eine ähnlich hohe Korrelation des Mechaniktests zur Abitur- und Mathematiknote, gefolgt von der Physiknote. Der Zusammenhang zur Deutschnote ist gering. Es ist auffällig, dass die Korrelationen zum Schulwissen höher als zu den beiden anderen Fachstufen sind. Dies ist plausibel, da dieses Wissen zu Beginn des Studiums wenig ausgeprägt ist und geraten werden muss. Ein vergleichbarer Zusammenhang der Abitur-, Mathematik- und Physiknote mit Fachwissenstests zeigt sich auch in anderen Studien (z. B. Schoppmeier et al., 2013), wobei der Zusammenhang dort höher ist. Momentan wird eine Konstruktvalidierung mit dem ähnlich konstruierten FW-Test von Woitkowski et al. (2011) durchgeführt.

Fachstufe	EAP/PV-Reliabilität	Varianz
Schulwissen	.83	1.43
Vertieftes Schulwissen	.78	1.05
Universitäres Wissen	.81	0.98

Tab. 1: EAP/PV-Reliabilitäten und Varianzen der Fachstufen beim FW

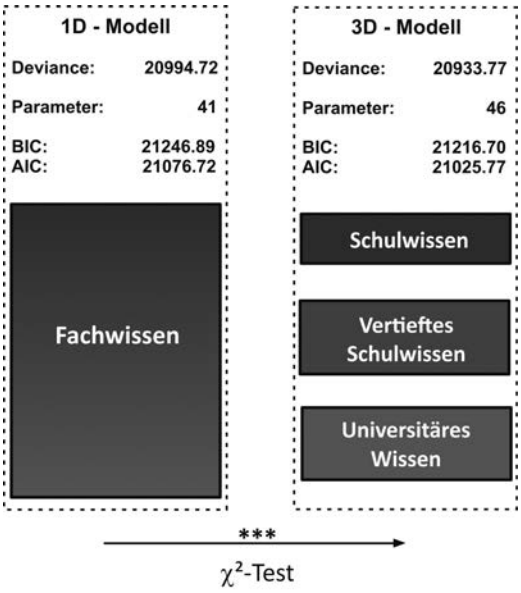


Abb. 5: Vergleich von 1D- und 3D-Modell bzgl. Fachstufen

	Schulwissen			Vertieftes Schulwissen			Universitäres Wissen		
	df	r	p	df	r	p	df	r	p
Rechentest	274	.67	< .001	274	.51	< .001	274	.41	< .001
Abiturnote	456	.46	< .001	456	.32	< .001	456	.36	< .001
Mathenote	447	.44	< .001	447	.27	< .001	447	.32	< .001
Physiknote	425	.35	< .001	425	.25	< .001	425	.26	< .001
Deutschnote	429	.15	.002	429	.10	.04	429	.12	.01

Tab. 2: Zusammenhang zwischen den drei Fachstufen und Kontrollvariablen

5.2 Fachdidaktisches Wissen

Zunächst wurden Analysen zur Inhaltsvalidierung durchgeführt, wobei die curriculare Validität als Teil der Inhaltsvalidität betrachtet wird (vgl. Jenßen et al., 2015, in diesem Beiheft). Hierzu wurden Curriculum-Analysen durchgeführt, indem die Lehrenden an den Erhebungsstandorten zur Passung der Testinhalte befragt wurden. Mithilfe dieser Daten konnte jedem Item ein Wert für die „curriculare Passung" zugewiesen werden. Auf einer Skala von 1 (sehr gute Passung) bis 5 (gar keine Passung) erzielten die Items Werte von 1.0 bis 3.2. Der Mittelwert aller Items lag bei 2.1 (Standardabweichung 0.6).

Zur weiteren Inhaltsvalidierung wurde eine „Think-aloud"-Studie mit 15 Studierenden (5 Lehramt HR, 5 Lehramt GyGe, 5 Fachwissenschaftler) durchgeführt. Sie sollte Rückschlüsse auf die Wissensbestände geben, die zur erfolgreichen Bearbeitung der Testitems nötig sind. Sofern beispielsweise ermittelt wurde, dass ein Item ausschließlich unter Nutzung von physikalischem FW beantwortet werden konnte, wurde es von der weiteren Verwendung ausgeschlossen. Schließlich wurden im Zusammenhang mit der Inhaltsvalidierung drei Experten (Physikdidaktiker) gebeten, die jeweiligen Testitems in das zugrunde liegende Modell (4 fachdidaktische Facetten und 3 kognitive Anforderungen) einzuordnen, um sicherzustellen, dass die Items als prototypisch für den intendierten Anforderungsbereich angesehen werden können. Dabei zeigt sich eine gute Übereinstimmung mit Cohens Kappa von .83 bis .89 für die Zuordnung der Items zu den Facetten und Kappa von .63 bis .81 für die Zuordnung zu den kognitiven Anforderungen.

Um weiter im Sinne einer Konstruktvalidierung die hypothetisch angenommene Struktur des FDW-Modells (vgl. 3.2) mit Methoden der probabilistischen Testtheorie zu prüfen, wurden 216 Lehramtsstudierende der Physik (davon 127 Gymnasialbereich und 71 Haupt-Realschulbereich) aus 12 Standorten getestet. Dabei wurde ein 1D-Rasch-Modell (EAP/PV-Reliabilität: .84; Varianz: .39) mit einem 4D-Rasch-Modell verglichen, das sich gemäß dem zugrunde liegenden Modell in die fachdidaktischen Facetten *Instruktionsstrategien, Schülervorstellungen, Experimente und Vermittlung eines angemessenen Wissenschaftsverständnisses* sowie *Fachdidaktische Konzepte* aufspaltet. Dabei passt das 4D-Modell hochsignifikant besser als das 1D-Modell (Chi-Quadrat-Test, vgl. Abb. 6). Wie beim 1D-Modell sind die Varianzen der einzelnen Facetten gering und ihre EAP/PV-Reliabilitäten sind akzeptabel (vgl. Tab. 3).

Eine Trennung der vier theoretisch angenommenen Facetten ist somit auch statistisch vertretbar, was als Indiz für die Konstruktvalidität gewertet werden kann. Im Zusammenhang dieser Analysen wurden Items mit schlechtem Fit aus dem Test entfernt (falls nicht $0.8 <$ gewichteter MNSQ < 1.2 und $-1.9 < T < 1.9$). Letztlich wurden unter Nutzung aller Erkenntnisse aus den berichteten Untersuchungen 43 von 91 Items für die weiteren Erhebungen zugelassen, womit gleichzeitig die Testzeit auf 60 Minuten begrenzt werden konnte.

Schließlich wurden theoretisch angenommene Zusammenhänge des mit dem FDW-Test ermittelten Konstrukts zu anderen Variablen überprüft (Prüfung eines nomologischen Netzes; vgl. Jenßen et al., 2015, in diesem Beiheft). Konkret wird erwartet, dass

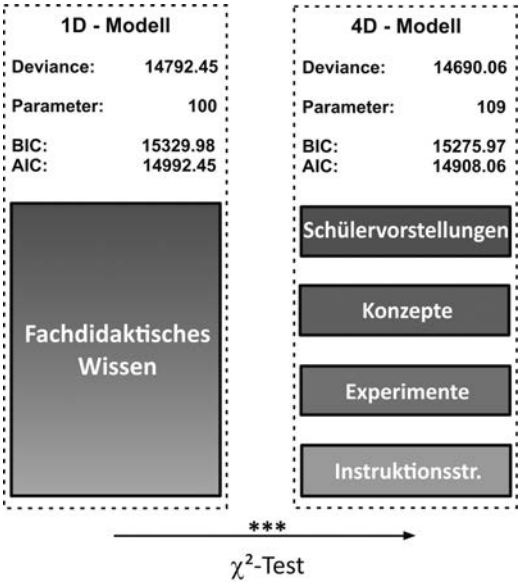


Abb. 6: Vergleich von 1D- und 4D-Modell fachdidaktischer Facetten

Facette	EAP/PV-Reliabilität	Varianz
Schülervorstellungen	.69	.52
Fachdidaktische Konzepte	.76	.78
Experimente	.74	.48
Instruktionsstrategien	.62	.49

Tab. 3: EAP/PV-Reliabilitäten und Varianzen der Facetten beim FDW

die Fähigkeitswerte des 4D-Rasch-Modells eine hohe Korrelation mit der Anzahl der Credit Points (CP) aufweisen, die in physikdidaktischen Seminaren erworben wurden (Indikator für die bisherigen Lerngelegenheiten), während geringere Korrelationen mit der Abiturnote (Indikator für allgemeine kognitive Leistungsfähigkeit), der letzten Physiknote (Indikator für das fachbezogene Vorwissen) und der letzten Deutschnote (Indikator für sprachliche Fähigkeiten) erwartet werden. Wie Tabelle 4 zeigt, sind erwartungskonform die höchsten und durchweg signifikanten Korrelationen mit den erworbenen CP in Physikdidaktik zu beobachten, was darauf hindeutet, dass der FDW-Test Wissenselemente erfasst, die tatsächlich in fachdidaktischen Seminaren erworben werden. Die restlichen Korrelationen sind geringer, aber unterschiedlich für die einzelnen Facetten, was wiederum die Annahme stützt, dass die vier Facetten unterschiedliche Konstrukte darstellen. Insgesamt können die Korrelationsanalysen als weiteres Argument für die

	Schüler-vorstellungen			Fachdidaktische Konzepte			Experimente			Instruktions-strategien		
	df	r	p	df	r	p	df	r	p	df	r	p
CP Physikdidaktik	194	.26	< .001	194	.20	.005	194	.36	< .001	194	.27	< .001
Abiturnote	205	.16	.020	205	.10	n. s.	205	.25	< .001	205	.12	n. s.
Physiknote	200	.06	n. s.	200	.05	n. s.	200	.24	< .001	200	.15	.032
Deutschnote	201	.07	n. s.	201	.12	n. s.	201	.20	.004	201	.12	n. s.

Tab. 4: Zusammenhang zwischen den vier Facetten und Kontrollvariablen

Konstruktvalidierung verstanden werden, da sie Testwertinterpretationen stützen, die auf die Erklärung von Leistungen abzielen.

5.3 Erklärungswissen

Um das EW in seiner Ausprägung auswerten zu können, braucht es einen validen Gütemaßstab. Zunächst wurden deshalb Einblicke in die Inhaltsvalidität gewonnen, anschließend wurde auf Basis der Kategorien ein solcher Maßstab entwickelt und mit einer Expertenerhebung validiert – in diesem Falle ist die Expertenbefragung zur Diskussion von Kriteriumsvalidität genutzt worden (vgl. Jenßen et al., 2015, in diesem Beiheft).

Die curriculare Validität der zu den Erklärungen ausgewählten Themen wurde unter anderem durch den Vergleich der Themen mit den Inhalten von gängigen Lehrbüchern der Experimentalphysik sowie von verbreiteten Physikschulbüchern sichergestellt. Alle Themen bilden dort Schwerpunkte.

Das diagnostische Verfahren selbst entspricht einer berufsbezogenen Anforderung, nämlich dem Erklären – dies spricht für Inhaltsvalidität („triviale Validität“; vgl. Jenßen et al., 2015, in diesem Beiheft). In einer begleitenden Interviewstudie wurden zusätzlich sieben Testpersonen zu den Erklärungen befragt. Es wurden besonders solche Testpersonen ausgewählt, die nicht auf die vorhergesehene Art auf die standardisierten Schülerfragen geantwortet haben. Die Probanden gaben auf direkte Nachfrage in Nachinterviews einheitlich an, dass sie die Settings als authentische Erklärungssituationen empfunden haben. Insbesondere haben sie nicht bemerkt, dass die Schülerinnen und Schüler trainiert waren. Zudem wurden die Testpersonen anhand ihrer Erklärungsvideos zu den Gründen ihrer von Ratern als problematisch eingeschätzten Reaktionen befragt (Video-Stimulated Recall). Aus diesen Interviews ging hervor, dass es sich bei den Gründen für das möglicherweise problematische Verhalten in den Erklärungen nicht um fachlich-physikalische Defizite, sondern um Probleme mit der Art des Erklärens handelte. Beispielsweise gaben zwei der Testpersonen an, dass sie Probleme mit der Erkundung des Vorwissens der Schüler hatten. Keine der Testpersonen nannte den fachlichen

Schwierigkeitsgrad als Grund. Die Probleme liegen also im intendierten Zielbereich, dies kann als weiteres Argument für Inhaltsvalidität gesehen werden.

Zur Vorbereitung der Entwicklung eines quantitativen Gütemaßes aus den qualitativen Daten für die Erklärungen wurde ein Expertenrating durchgeführt, das der konkurrenten Validierung dienen soll (Kriteriumsvalidität). Das naheliegende Außenkriterium zu erheben – nämlich das Verständnis der Schülerinnen und Schüler, die als Adressaten fungieren – konnte hier nicht angewendet werden. Die Schülerinnen und Schüler sind schließlich im Sinne der Methode zu standardisiertem Verhalten trainiert worden und kannten die Themen. Deshalb wurden sieben Experten (Fachdidaktiker und Lehrkräfte mit hoher Lehrerfahrung) jeweils Paare von insgesamt 16 Erklärungsvideos gezeigt. Die Experten mussten hochinferent entscheiden, in welchem Video die bessere Erklärung vorlag. Jedes Paar wurde von fünf Experten beurteilt. Diese Validierungsstrategie durch Paarvergleiche wurde bei komplexen Merkmalen schon häufiger gewählt (Kulgemeyer & Schecker, 2013), gerade bei hoher Interraterreliabilität mehrerer Rater wird ihr Aussagekraft zugemessen. Es ergab sich hier eine sehr gute Beurteilerübereinstimmung über die bessere Erklärung (95 %, Fleiss' $\kappa = .80$). Diese Expertenentscheidung kann zur späteren konvergenten Validierung eines Bewertungsmaßes für Erklärungswissen verwendet werden.

Anschließend wurde schrittweise ein Bewertungsmaßstab für EW entwickelt. Aus den bis dahin vorliegenden 45 Kategorien bzgl. guter Erklärungen wurden diejenigen ausgewählt, die nach der Theorie positiv für Erklärungsqualität sind (z. B. „Evaluation des Verständnisses", siehe 3.3). Ein weiteres Auswahlkriterium war das gehäufte Auftreten in solchen Erklärungen, die von den Experten als besonders gut oder als besonders schlecht eingeschätzt wurden. Die Kategorien, die in diesen beiden „Polen" von Erklärungen besonders häufig vorkommen, wurden für die weiteren Analysen berücksichtigt. Über das Vorkommen der verbleibenden 26 positiven und negativen Kategorien in einer Erklärung wurde analog zu dichotomen Testitems summiert (0/1), wobei die negativen Kategorien umgekehrt gepolt berücksichtigt wurden. Es ergibt sich eine Skala für EW mit zufriedenstellender Reliabilität (Cronbachs $\alpha = .74$).

In nächsten Schritt sollte unter dem Aspekt der Auswertungsökonomie die Anzahl der Kategorien reduziert werden, ohne Validität oder Reliabilität der Skala zu verringern. Dazu wurden in einem Wechselspiel der Prüfung der Auswirkungen auf Reliabilität (Item-Skala-Korrelation, Änderungen in Cronbachs Alpha) und Prognosefähigkeit (Übereinstimmung mit dem Expertenrating, Cohens Kappa) schrittweise Kategorien gestrichen. Es sollten dabei Kategorien gestrichen werden, ohne diese beiden Kriterien zu verschlechtern. Die verbleibenden zwölf Kategorien sagen die Expertenentscheidung über die bessere Erklärung im Paarvergleich zweier Videomittschnitte voraus. Jeder Bereich des Kommunikationsmodells (Sprache, Beispiele, Darstellungsform, Mathematisierung) ist dabei im Maß mit mindestens einer Kategorie vertreten (z. B. „Gibt Zahlenbeispiele" für Mathematisierung). Die Übereinstimmung zwischen den mit diesem schlankeren Maß getroffenen Entscheidungen über die bessere Erklärung im Paarvergleich und den Ergebnissen aus dem Expertenrating beträgt Cohens $\kappa = .75$ (87.5 %). Die Reliabilität der Skala steigt sogar (Cronbachs $\alpha = .77$) und ist damit im guten Bereich.

6. Diskussion und Ausblick

Im Projektverbund ProfiLe-P wurde ein Ansatz zur Messung des Professionswissens von Physiklehrkräften entwickelt und validiert, der zentrale Wissenselemente in einem gemeinsamen Rahmenmodell differenziert abbildet und damit bereichsübergreifend überprüfbar macht. Ausgehend von drei bereichsspezifischen Kompetenzmodellen können so physikalisches Fachwissen (FW), physikbezogenes fachdidaktisches Wissen (FDW) und die Fähigkeit zum Erklären von Physik (EW) zueinander in Beziehung gesetzt werden. Damit können erstmals im Rahmen von Large-Scale-Assessments in den Naturwissenschaften Zusammenhänge zwischen dem Professionswissen und Elementen des Handelns von angehenden Lehrkräften verhältnismäßig ökonomisch untersucht werden, indem zeitlich begrenzte (10 Min.) Lehr-Lern-Rollenspiele mit trainierten „Schülerinnen und Schülern“ in Laborsituationen videografiert werden.

Wie für alle empirischen Studien stellt sich die Frage, inwieweit dem fundamentalen und zugleich komplexesten Gütekriterium, der Validität, Rechnung getragen wird. Dementsprechend wurde in ProfiLe-P eine Reihe von Untersuchungen durchgeführt, um Argumente zur Inhaltsvalidität, zur Konstruktvalidität und zur Kriteriumsvalidität zu erhalten. Insgesamt deuten die durchgeführten Studien in der Summe darauf hin, dass die gewählten Testverfahren valide Testwertinterpretationen im Hinblick auf die Erklärung von Leistungen zulassen. Bislang liegt jedoch noch keine ausreichende Evidenz vor, inwieweit die Tests Interpretationen zulassen, die auf eine (individuelle) Bewertung von Leistung abzielen, wenn die Leistungen einer Person mit den Leistungen einer anderen Person verglichen werden (vgl. Hartig et al., 2012).

Darüber hinaus bleiben für die laufenden Erhebungen noch einige offene Fragen. So werden Analysen zur Trennbarkeit des vertieften Schulwissens vom universitären Wissen möglicherweise dadurch verzerrt, dass nicht alle spezifischen Fähigkeiten im FW-Test berücksichtigt werden. Die Facetten des dem FDW-Test zugrunde liegenden Modells wiederum lassen sich zwar empirisch trennen und ermöglichen so eine differenziertere Messung als vergleichbare Tests bisheriger Studien, allerdings ist der Test insgesamt etwas zu schwer. Damit sind Aussagen zu leistungsschwachen Gruppen nur bedingt möglich. Weiterhin findet die fachliche Korrektheit der Erklärung beim Maß für das EW wenig Beachtung, bislang wird vor allem die Adressatengemäßheit operationalisiert. Daran wird derzeit gearbeitet, um den dialogischen Aspekt des Erklärens stärker hervorzuheben und differenziertere Aussagen zu ermöglichen. Wenn die angestrebte Stichprobe von N = 200 erreicht ist, könnte auch im EW mit Methoden der probabilistischen Testtheorie eine Modellierung geprüft werden. Wir gehen davon aus, dass die Daten aus den laufenden Erhebungen weitere Aufschlüsse ermöglichen. Zudem hat sich die Kodierung der offenen Items des FDW-Tests als sehr aufwendig erwiesen, da nicht immer eindeutig zu entscheiden ist, ob eine Antwort als falsch oder richtig zu beurteilen ist. Auch die Kodierung der videografierten Erklärungen des Tests zur Erfassung des EW ist bislang noch sehr aufwendig, sodass hier für Folgestudien zu prüfen ist, inwieweit einzelne Anforderungen mit geschlossenen oder online-basierten Items erfasst werden können.

Grundsätzlich ist das der Itementwicklung zugrunde liegende Modell konzeptionell anschlussfähig an die aktuelle Diskussion zur Lehrerbildungsforschung, da es eine Weiterentwicklung bzw. Adaption von vorliegenden Modellen darstellt. Daher werden auf Basis der Auswertung der Haupterhebung, in der alle Instrumente in einer gemeinsamen Stichprobe zum Einsatz kommen, vergleichende Auswertungen und Interpretationen mit Studien in anderen Fachdomänen, bei denen Bezüge zwischen Bereichen des Professionswissens einerseits und Aspekten der Unterrichtsqualität oder des Lehrerhandelns in realitätsnahen Situationen andererseits hergestellt werden (für Mathematiklehrkräfte z. B. Kunter et al., 2013), grundsätzlich ermöglicht. Einschränkend muss dabei angemerkt werden, dass sich die Ergebnisse der Haupterhebung nur auf die Erhebungen an vier Lehrerbildungsstandorten in drei Bundesländern beziehen (werden), was die Generalisierbarkeit der Ergebnisse zum Teil begrenzt. Dies ist vor allem der Tatsache geschuldet, dass die Probanden zeitlich sehr stark in Anspruch genommen werden (mehrere Testhefte an den jeweiligen Messzeitpunkten), was die Probandenakquise an „fremden" Lehrerbildungsstandorten deutlich erschwert. Auf der anderen Seite wurden Pilotierungen, Begleituntersuchungen und curriculare Analysen für einzelne Testteile an insgesamt zwölf Lehrerbildungsstandorten durchgeführt und es wurden allgemeine Standards der Lehrerbildung für die Entwicklung der Tests herangezogen, sodass zumindest im Zusammenhang mit der Testentwicklung bestmöglich gewährleistet sein sollte, dass die Tests bundesweit einsetzbar sind.

Schließlich verfolgen wir u. a. die folgenden bereichsverbindenden Fragestellungen im Zusammenhang mit der noch laufenden Haupterhebung, bei der alle Instrumente in einer gemeinsamen Stichprobe eingesetzt werden:

- Wie viel Varianz im Erklärungswissen lässt sich durch das Fachwissen und das fachdidaktische Wissen aufklären?
- Gibt es über globale Zusammenhänge zwischen den untersuchten drei Wissensbereichen hinausgehende spezifische Zusammenhänge auf der Ebene jeweiliger Unterfacetten?
- Wie entwickeln sich die verschiedenen Teilbereiche des Professionswissens im Längsschnitt des ersten und zweiten Studienjahres?

Perspektivisch bieten die entwickelten Testinstrumente die Möglichkeit, Kompetenzstandserhebung in zentralen Abschnitten der universitären Lehramtsausbildung durchführen zu können, womit Voraussetzungen sowohl für die Durchführung von Interventionsstudien zur Wirkung in der fachlichen und fachdidaktischen Ausbildung als auch für die Evaluation von Studienprogrammen im Lehramtsstudium Physik gegeben sind. Dabei erscheint ein Transfer der genutzten Methoden im Zusammenhang mit der Entwicklung des Kompetenzmodells und der Testinstrumente in andere Fächer grundsätzlich möglich.

Literatur

Abell, S. K. (2007). Research on Science Teacher Knowledge. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of research on science education* (S. 1105–1149). Mahwah: Lawrence Erlbaum.

Anderson, L. W., & Krathwohl, D. R. (Hrsg.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

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

Blömeke, S., & Zlatkin-Troitschanskaia, O. (2013). *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs* (KoKoHs Working Papers, 1). Berlin/Mainz: Humboldt-Universität/Johannes Gutenberg-Universität.

Borowski, A., & Riese, J. (2010). Physikalisch-fachdidaktisches Wissen – Was kommt in der Praxis an? *Praxis der Naturwissenschaften – Physik in der Schule, 59*(5), 5–8.

Brown, G. (2006). Explaining. In O. Hargie (Hrsg.), *The handbook of communication skills* (S. 195–228). East Sussex: Taylor & Francis.

Deutsche Physikalische Gesellschaft (DPG) (2014). *Zur fachlichen und fachdidaktischen Ausbildung für das Lehramt Physik*. Bad Honnef: Deutsche Physikalische Gesellschaft.

Fischer, H. E., Borowski, A., & Tepner, O. (2012). Professional Knowledge of Science Teachers. In B. J. Fraser, K. G. Tobin & C. J. McRobbie (Hrsg.), *Second international handbook of science education* (S. 435–448). Dordrecht: Springer.

Geelan, D. (2012). Teacher Explanations. In B. Fraser, K. Tobin & C. McRobbie (Hrsg.), *Second International Handbook of Science Education* (S. 987–999). Dordrecht: Springer.

Gesellschaft für Fachdidaktik (GFD) (2004). *Kerncurriculum Fachdidaktik. Orientierungsrahmen für alle Fachdidaktiken*. Entwurf des Arbeitskreises „Kerncurriculum" Fachdidaktik der Gesellschaft für Fachdidaktik e. V.

Gramzow, Y., Riese, J., & Reinhold, P. (2013). Modellierung fachdidaktischen Wissens angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften, 19*, 31–49.

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143–173). Berlin: Springer.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141–158.

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik, 61. Beiheft*, 11–31.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73.

Korneck, F., Lamprecht, J., Wodzinski, R., & Schecker, H. (2010). *Quereinsteiger in das Lehramt Physik – Lage und Perspektiven der Physiklehrerausbildung in Deutschland*. Bad Honnef: Deutsche Physikalische Gesellschaft.

Kröger, J., Neumann, K., & Petersen, S. (2013). Messung professioneller Kompetenz im Fach Physik. In S. Bernholt (Hrsg.), *Inquiry-based Learning – Forschendes Lernen.* Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012 (S. 533–535). Kiel: IPN.

Kulgemeyer, Ch., & Schecker, H. (2013). Students explaining science – assessment of science communication competence. *Research in Science Education, 43*, 2235–2256.

Kultusministerkonferenz (KMK) (2008). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerinnen- und Lehrerbildung* (Beschluss der Kultusministerkonferenz vom 16. Oktober 2008). Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., et al. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820.

Mayring, P. (2003). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Weinheim/Basel: Beltz.

Moosbrugger, H., & Kelava, A. (Hrsg.) (2012). *Testtheorie und Fragebogenkonstruktion* (2. Aufl.). Berlin: Springer.

Ohle, A. (2010). *Primary school teachers' content knowledge in physics and its impact on teaching and students' achievement*. Berlin: Logos.

Olszewski, J. (2010). *The impact of physics teachers' Pedagogical content knowledge on teacher action and student outcomes*. Berlin: Logos.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Hrsg.) (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe*. Münster: Waxmann.

Park, S., & Chen, Y. C. (2012). Mapping out the integration of the components of pedagogical content knowledge (PCK) – Examples from high school biology classrooms. *Journal of Research in Science Teaching, 49*(7), 922–941.

Riese, J., & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen – Empirische Hinweise für eine Verbesserung des Lehramtsstudiums. *Zeitschrift für Erziehungswissenschaft, 15*(1), 111–143.

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

Schecker, H. (1985). *Das Schülervorverständnis zur Mechanik – eine Untersuchung in der Sekundarstufe II unter Einbeziehung historischer und wissenschaftstheoretischer Aspekte* (Dissertation). Bremen: Universität.

Schoppmeier, F., Borowski, A., & Fischer, H. E. (2013). Validierung eines Kompetenzmodells für Physik in der Sekundarstufe II. In S. Bernholt (Hrsg.), *Inquiry-based Learning – Forschendes Lernen*. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hannover 2012 (S. 206–208). Kiel: IPN.

Shulman, L. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*, 4–14.

Stichweh, R. (1996). Professionen in einer funktional differenzierten Gesellschaft. In A. Combe & W. Helsper (Hrsg.), *Pädagogische Professionalität – Untersuchungen zum Typus pädagogischen Handelns* (S. 49–69). Frankfurt a. M.: Suhrkamp.

Tepner, O., Borowski, A., Dollny, S., Fischer, H. E., Jüttner, M., et al. (2012). Modell zur Entwicklung von Testitems zur Erfassung des Professionswissens von Lehrkräften in den Naturwissenschaften. *Zeitschrift für Didaktik der Naturwissenschaften, 18*, 7–28.

Vogelsang, C., & Reinhold, P. (2013). Zur Handlungsvalidität von Tests zum professionellen Wissen von Lehrkräften. *Zeitschrift für Didaktik der Naturwissenschaften, 19*, 103–128.

Wellenreuther, M. (2005). *Lehren und Lernen – aber wie? Empirisch-experimentelle Forschung zum Lehren und Lernen im Unterricht*. Baltmannsweiler: Schneider.

Woitkowski, D., Riese, J., & Reinhold, P. (2011). Modellierung fachwissenschaftlicher Kompetenz angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften, 17*(1), 289–313.

Zlatkin-Troitschanskaia, O., & Kuhn, C. (2010). Messung akademisch vermittelter Fertigkeiten und Kenntnisse von Studierenden bzw. Hochschulabsolventen – Analysen zum Forschungsstand. In *Arbeitspapiere WP Heft 56*. http://www.wipaed.uni-mainz.de/ls/382.php [11.04.2014].

Abstract: For improving teacher education, there has been an increasing interest in describing teachers' professional competencies and their development in the course of implementing educational programs. The focus of the present project is on modeling and measuring domain-specific and generic competencies that future physics teachers acquire during their university studies. The model comprises characteristics and relationships between physics content knowledge, pedagogical content knowledge, and skills for explaining physics phenomena. Based on the model, assessment instruments were developed and applied as paper-and-pencil-tests and videotaped expert-novice dialogues for measuring the competencies in a large sample of physics student teachers. Trials and validation suggest that our instruments are valid in terms of content and construct validities.

Keywords: Professional Knowledge, Student Teachers, Physics, Competencies, Validity

Anschrift der Autor(inn)en

Dr. Josef Riese, Universität Paderborn, AG Didaktik der Physik,
Warburger Straße 100, 33098 Paderborn, Deutschland
E-Mail: josef.riese@upb.de

Dr. Christoph Kulgemeyer, Universität Bremen, Institut für Didaktik der Naturwissenschaften,
Abteilung Physikdidaktik, Otto-Hahn-Allee 1, 28359 Bremen, Deutschland
E-Mail: kulgemeyer@physik.uni-bremen.de

Simon Zander, Universität Duisburg-Essen, Physikdidaktik,
Universitätsstraße 2, 45141 Essen, Deutschland
E-Mail: simon.zander@uni-due.de

Prof. Dr. Andreas Borowski, Universität Potsdam, AG Didaktik der Physik,
Karl-Liebknecht Straße 24/25, 14476 Potsdam, Deutschland
E-Mail: andreas.borowski@uni-potsdam.de

Prof. Dr. Hans E. Fischer, Universität Duisburg-Essen, Lehrstuhl für Didaktik
der Physik, Universitätsstraße 2, 45141 Essen, Deutschland
E-Mail: hans.fischer@uni-due.de

Yvonne Gramzow, Universität Paderborn, AG Didaktik der Physik,
Warburger Straße 100, 33098 Paderborn, Deutschland
E-Mail: ygramzow@upb.de

Prof. Dr. Peter Reinhold, Universität Paderborn, AG Didaktik der Physik,
Warburger Straße 100, 33098 Paderborn, Deutschland
E-Mail: peter.reinhold@upb.de

Prof. Dr. Horst Schecker, Universität Bremen, Institut für Didaktik der Naturwissenschaften,
Abteilung Physikdidaktik, Otto-Hahn-Allee 1, 28359 Bremen, Deutschland
E-Mail: schecker@physik.uni-bremen.de

Elisabeth Tomczyszyn, Universität Bremen, Institut für Didaktik der Naturwissenschaften,
Abteilung Physikdidaktik, Otto-Hahn-Allee 1, 28359 Bremen, Deutschland
E-Mail: e.tomczyszyn@uni-bremen.de

Simone Dunekacke/Lars Jenßen/Sigrid Blömeke

Mathematikdidaktische Kompetenz von Erzieherinnen und Erziehern

Validierung des KomMa-Leistungstests durch die videogestützte Erhebung von Performanz

Zusammenfassung: In diesem Beitrag wird die prognostische Validität eines Papier-und-Bleistift-Tests zur Erfassung des mathematikdidaktischen Wissens von angehenden Erzieherinnen und Erzieherin untersucht, der im Projekt KomMa entwickelt wurde. Aus der Lehrerbildungsforschung ist bekannt, dass mathematikdidaktisches Wissen ein Prädiktor für die Unterrichtswahrnehmung und Performanz von jungen Lehrkräften ist. Entsprechende Hypothesen wurden für Erzieherinnen und Erzieher aufgestellt. Die Wahrnehmung von Kindergartensituationen und die Performanz in Form von Handlungsplänen wurden mit einem videobasierten Assessment erhoben und zusammen mit dem Wissenstest bei 354 angehenden Erzieherinnen und Erziehern eingesetzt. Mit der Prüfung konkurrierender Strukturgleichungsmodelle kann gezeigt werden, dass das mathematikdidaktische Wissen einen signifikanten direkten Einfluss auf die Situationswahrnehmung und einen signifikanten indirekten Einfluss auf die Performanz der Erzieherinnen und Erzieher hat. Damit liegen nicht nur Hinweise für die prognostische Validität des entwickelten Leistungstests vor, sondern auch der Zusammenhang von Wissen, Wahrnehmung und Handlungsplanung wird deutlich.

Schlagworte: Erzieher/innen, mathematikdidaktisches Wissen, Validierung, Leistungstest, videobasierte Kompetenzerfassung

Frühe mathematische Bildung ist ein wichtiger Prädiktor für Schulerfolg (Krajewski & Schneider, 2009; Duncan et al., 2007). Ihr Gelingen hängt allerdings von einer anregungsreichen Lernumgebung und deren pädagogisch-didaktischer Begleitung ab (van Oers, 2009; Reynolds, 1995; Klibanoff, Levine, Huttenlocher, Vasilyeva & Hedges, 2006). Aufgrund der historischen Entwicklung von Kindertagesstätten als Betreuungs- und Erziehungseinrichtungen wurde auf eine fachbezogene professionelle Kompetenz der Erzieher/innen lange Zeit wenig Wert gelegt, was sich auch in einem Mangel an empirischer Forschung zu ihrer professionellen Kompetenz zeigt. Dies gilt insbesondere für den mathematischen Bereich (Thole, 2010; Fried & Roux, 2009; National Advisory Panel, 2008). Eine Analyse der Lehrpläne für Fachschulen, an denen in Deutschland die Mehrheit der Erzieher/innen ausgebildet wird, hat entsprechend gezeigt, dass Mathematik und Mathematikdidaktik nur in wenigen Ausnahmen zum Kerncurriculum gehören. Das Projekt KomMa[1] leistet einen Beitrag zur Schließung dieser Forschungslücke, in-

1 KomMa – *Struktur, Niveau und Entwicklung professioneller Kompetenz von Erzieher/innen im Bereich Mathematik* – ist ein Kooperationsprojekt der Humboldt-Universität zu Berlin und

dem es die professionelle Kompetenz von angehenden Erzieher/innen untersucht. Ziel des vorliegenden Beitrags ist in diesem Zusammenhang der Nachweis prognostischer Validität des mathematikdidaktischen Leistungstests, dass er also in der Lage ist, handlungsrelevantes Professionswissen zu erfassen.

1. Theoretischer Hintergrund

1.1 Prognostische Validität: Gelingt die Vorhersage von erfolgreichem beruflichem Handeln?

Die Einhaltung von Gütekriterien gehört zu den wissenschaftlichen Standards in der empirischen Bildungsforschung (Bortz & Döring, 2006; Rost, 2004). Zu den drei Hauptgütekriterien von empirischen Messungen zählen Objektivität, Reliabilität und Validität. Die Validität gilt als zentrales, aber komplexes Gütekriterium, welches gleichzeitig schwierig nachzuweisen ist (Jenßen, Dunekacke & Blömeke, 2015, in diesem Beiheft). Im Mittelpunkt dieses Beitrages steht die Kriteriumsvalidität, bei der es um die Frage geht, inwieweit von den Testwerten auf Situationen außerhalb des Tests geschlossen werden kann (sog. diagnostische Entscheidungen; Hartig, Frey & Jude, 2012, S. 155). Es wird dabei unterschieden zwischen der Übereinstimmungsvalidität, bei der das Kriterium parallel zur Testsituation liegt, und der prognostischen Validität, bei der das Kriterium in der Zukunft liegt (S. 157). Diagnostische Entscheidungen sind besonders für die pädagogisch-psychologische Praxis relevant (S. 155), haben ihre Bedeutung aber auch im Kontext der Forschung. Für die Bildungsforschung bedeutet dies beispielsweise zu berücksichtigen, dass Wissen im Kontext von Ausbildung und Studium erworben wird, um berufliche Anforderungen erfolgreich zu bewältigen. Ein Test kann damit als prognostisch valide angesehen werden, wenn er deren Bewältigung – als das Kriterium – erfolgreich vorhersagen kann (Cronbach & Meehl, 1955).

1.2 Mathematikdidaktisches Wissen von Erzieher/innen

Studien zum professionellen Wissen von angehenden Erzieher/innen gibt es national und international kaum (Thole, 2010; Fried & Roux, 2009; National Advisory Panel, 2008). Ein anderes Bild zeichnet sich in der Forschung zur Lehrerausbildung ab. Bereits seit Mitte der 1980er-Jahre gibt es Studien, die sich mit der Struktur des professionellen Wissens befassen. Viele davon beziehen sich auf die Arbeiten von Shulman (1986), der darlegt, dass professionelles Wissen von Lehrkräften aus unterschiedlichen Wissensfacetten besteht, welche von motivational-affektiven Facetten flankiert werden. Im Be-

der Alice-Salomon-Hochschule Berlin und wird vom Bundesministerium für Bildung und Forschung (FKZ: 01PK11002A) im Rahmen der Förderinitiative KoKoHs – *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor* – gefördert.

reich Mathematik handelt es sich um das mathematische Fachwissen, das mathematikdidaktische Wissen und das allgemein-pädagogische Wissen (ebd.).

Diese heuristische Unterscheidung konnte teilweise weiter ausdifferenziert und in verschiedenen empirischen Studien zum professionellen Wissen von Lehrkräften bestätigt werden (z. B. Ball & Bass, 2009; Blömeke, Felbrich, Müller, Kaiser & Lehmann, 2008). Für das Niveau an mathematikdidaktischem Wissen angehender Primarstufenlehrkräfte in Deutschland zeigten sich dabei deutliche Unterschiede in Abhängigkeit der Ausbildungsform (Blömeke, Kaiser, Döhrmann, Suhl & Lehmann, 2010, S. 239). Relativ gute Werte erreichten Studierende, die ein reines Primarstufenlehramt mit dem Schwerpunkt Mathematik studiert hatten. Ohne Mathematik als Schwerpunkt konnten dagegen vielfach nur schwache Leistungen gezeigt werden (ebd.). Validierungsstudien machten anschließend deutlich, dass diese Unterschiede direkte Auswirkungen auf die handlungsbezogenen Fähigkeiten der Lehrkräfte haben (Blömeke et al., im Druck).

An vergleichbaren Studien zum professionellen Wissen von angehenden Erzieher/innen im Bereich Mathematik mangelt es. In konzeptioneller Hinsicht stellen verschiedene Autoren allerdings heraus, dass das mathematikdidaktische Wissen vermutlich auch bei dieser Personengruppe – neben dem mathematischen Fachwissen und den Überzeugungen – zentraler Bestandteil der professionellen Kompetenz ist und eine Schlüsselrolle in Bezug auf ihr Handeln in Kindertageseinrichtungen einnimmt (Anders, 2012; Gasteiger, 2010, S. 153; Lee, 2010; Sarama & Clements, 2009, S. 354; Ginsburg & Ertle, 2008, S. 46). Erste empirische Belege hierfür konnte Lee (2010) vorlegen, der mathematikdidaktisches Wissen differenziert nach mathematischen Bildungsinhalten erfasst und für im Beruf stehende Erzieher/innen jeweils Niveauunterschiede in Abhängigkeit vom Ausbildungsabschluss feststellt. Erzieher/innen erreichen eine höhere Punktzahl, wenn es um zahlbezogenes mathematikdidaktisches Wissen geht, als im Bereich der Raumwahrnehmung (ebd.). Darüber hinaus erreichen solche mit einem Master- oder Doktorabschluss deutlich höhere Werte als jene, die über einen Bachelorabschluss verfügen (ebd.). Lee, Meadows und Lee (2003) konnten dann im Anschluss zeigen, dass Erzieher/innen mit einem höheren mathematikdidaktischen Wissen qualitativ bessere mathematische Lerngelegenheiten für Kinder schaffen.

Im Forschungsprojekt KomMa wurde vor diesem Hintergrund ein Leistungstest entwickelt, mit dem das mathematikdidaktische Wissen gemessen und Aussagen über dessen Zusammenhänge mit Handlungsmerkmalen ermöglicht werden sollen. Theoretische Grundlage war ein, ebenfalls im Rahmen von KomMa entwickeltes, Kompetenzstrukturmodell mit vier Subdimensionen: Wissen über die Gestaltung von situativen und geplanten mathematischen Bildungsprozessen in Kindertageseinrichtungen, Wissen über die mathematische Entwicklung von Kindern, Wissen zu deren Diagnostik und Wissen zu deren Förderung (Jenßen et al., im Druck). Die Subdimensionen und theoretisch antizipierten Inhalte gehen auf eine Analyse der Bildungspläne für Kindertageseinrichtungen in allen 16 Bundesländern und eine Analyse einschlägiger Fachliteratur zur frühen mathematischen Bildung zurück (ebd.). Im Bereich der mathematikbezogenen Diagnostik geht es beispielsweise darum, ob die Erzieher/innen in einer scheinbar falschen mathematischen Handlung eines Kindes (Zählen von Buntstiften) ein übergeordnetes *Mus-*

ter (Zählen nach Farbgruppen, z. B. alle roten und pinken Stifte) erkennen. Im Rahmen einer Befragung von Expert/innen konnte die Inhaltsvalidität der entwickelten Items nachgewiesen werden (Jenßen et al., 2015).

1.3 Verbindung von Wissen und Performanz

Weinert (2001) hat Kompetenz als mehrdimensionales Konstrukt mit kognitiven Fähigkeiten und Fertigkeiten beschrieben, das unter Berücksichtigung von motivational-affektiven Aspekten und unter unterschiedlichen situationalen Bedingungen zur erfolgreichen Bearbeitung von beruflichen Anforderungen führt. Eine lineare Umsetzung des Wissens in einer Handlungssituation bedeutet dies also nicht (Blömeke, König et al., 2014; Blömeke, Busse, Kaiser, König & Suhl, re-submitted after revisions), sondern die Umsetzung beruht auf einer Umstrukturierung und gezielten, adaptiven Anwendungen des Wissens in den jeweiligen Anforderungssituationen (Weinert, Schrader & Helmke, 1990). Blömeke, König et al. (2014) konnten in Bezug auf Mathematiklehrkräfte der Sekundarstufe aber zeigen, dass deren mathematikdidaktisches Wissen in der Phase des Berufseinstieges ein Prädiktor für die Wahrnehmung, Interpretation und Entscheidung über Handlungsstrategien in Unterrichtssituationen ist. Solche Zusammenhänge von Wissen und Performanz werden auch im theoretisch erarbeiteten Kompetenzmodell der Frühpädagogik angenommen (Abb. 1).

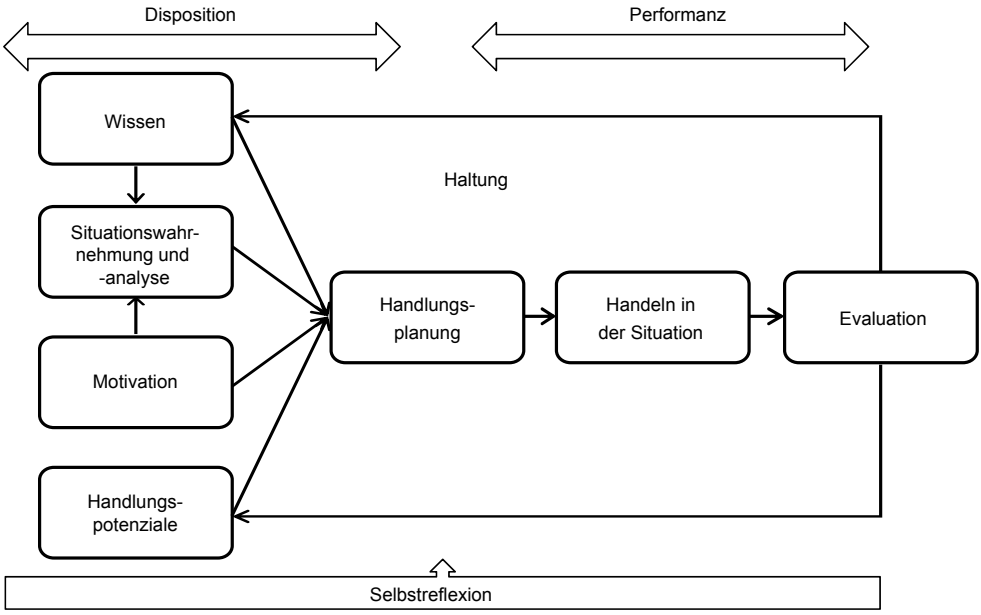


Abb. 1: Kompetenzmodell der Frühpädagogik (in Anlehnung an Fröhlich-Gildhoff, Nentwig-Gesemann & Pietsch, 2011, S. 17)

Die Handlungsfähigkeit besteht danach aus dem „wechselseitigen Zusammenspiel“ (Fröhlich-Gildhoff, Nentwig-Gesemann & Pietsch, 2011, S. 18) von Wissen, Handlungspotenzialen, Motivation und Situationswahrnehmung. Ein Teilausschnitt des Modells ist Grundlage dieser Studie. Dabei werden das mathematikdidaktische Wissen und die Situationswahrnehmung als zwei bedeutsame Merkmale der professionellen Kompetenz von Erzieher/innen erfasst und die Handlungsplanung als Indikator ihrer Performanz in Kindergartensituationen.

1.4 Situationswahrnehmung

Situationswahrnehmung wird häufig als eine wichtige Fähigkeit von Lehrkräften sowie Erzieher/innen konzeptualisiert. Dies ist darauf zurückzuführen, dass pädagogische Situationen hochkomplex und wenig standardisiert sind (Fröhlich-Gildhoff et al., 2011, S. 17). Um adaptiv handeln zu können, ist das Wahrnehmen der spezifischen Situationsmerkmale zentrale Voraussetzung (Thonhauser, 2007; Star & Strickland, 2008; van Es & Sherin, 2006).

Van Es und Sherin (2006, S. 245) beschreiben Wahrnehmen als ersten Schritt des situationsspezifischen Handlungsprozesses von Lehrkräften, der aus drei Phasen besteht: *„(a) identifying what is important in a teaching situation; (b) using what one knows about the context to reason about a situation; and (c) making connections between specific events and broader principles of teaching and learning“*. Der erste Schritt wird dabei gerade in Bezug auf Auszubildende als zentral gesehen (Star & Strickland, 2008). Um zu bestimmen, wie die entsprechenden Fähigkeiten ausgeprägt sind, werden objektiv beschreibbare Merkmale einer Situation, wie beispielsweise das Thema, das Material oder der soziale Kontext, als Indikatoren herangezogen (Thonhauser, 2007). Die Qualität der Situationswahrnehmung hängt unter anderem von der Erfahrung der Erzieher/innen ab und entwickelt sich ständig weiter (Perrez, Huber & Geißler, 2001, S. 366; Star & Strickland, 2008).

Der Fokus der vorliegenden Studie liegt in diesem Bereich auf der Wahrnehmung mathematikdidaktischer Aspekte. Eine solch fachbezogene Wahrnehmung ist erst in den vergangenen Jahren Gegenstand der Forschung geworden (Blomberg, Stürmer & Seidel, 2011). Eine fachbezogene Wahrnehmung ist dadurch gekennzeichnet, dass nicht nur allgemein-pädagogische Aspekte der Lehr-Lernsituation erkannt werden, sondern dass inhaltliche Merkmale wie beispielsweise die (Fehl-)Vorstellungen der Lernenden erkannt werden (Sherin, 2007). In der vorliegenden Studie handelt es sich zum Beispiel um Wahrnehmungen der mathematischen Materialien (z. B. „Es werden verschiedenen Darstellungen der Zahl gezeigt“) oder der mathematischen Handlungen der Kinder (z. B. „Die Kinder versuchen ihre Größe mit der Hand anzuzeigen (Nutzen von ‚Strategien‘)“).

1.5 Handlungsplanung

Handeln wird erneut als ein Prozess verstanden, der sich aus verschiedenen Phasen zusammensetzt (Fröhlich-Gildhoff et al., 2011, S. 17; Widulle, 2009, S. 19; Gudjons, 2008, S. 46 ff.; Wild & Krapp, 2001, S. 519). Die Phasen lassen sich im Wesentlichen in eine Planungs- oder Vorbereitungsphase, die auch Entscheidungen über Handlungsziele beinhaltet, sowie eine Durchführungsphase und eine Evaluationsphase unterscheiden (ebd.).

Konzeptionell kann angenommen werden, dass die Handlungsplanung der Situationswahrnehmung folgt (Hogrebe, Schulz & Böttcher, 2012; Schäfer, 2005). Aufbauend auf einer theoretisch fundierten Situationswahrnehmung werden idealerweise verschiedene situationsspezifische Alternativen geprüft und eine wird ausgewählt (ebd.). Die weitere Umsetzung hängt dann auch von der spezifischen Situation ab (Joas, 1996, S. 236). Die Handlungsplanung ist demnach als Teil der Handlung zu verstehen (Fröhlich-Gildhoff et al., 2011), aber nicht zwingend identisch mit ihr (Hogrebe et al., 2012). Dennoch wird Handlungsplanung, bewusst oder unbewusst, für Erzieher/innen als bedeutsam angesehen, da sie die Aufmerksamkeit auf die Interessen des Kindes richtet (Schäfer, 2005).

In der hier präsentierten Studie stehen wie bei der Situationswahrnehmung spezifisch mathematikdidaktische Handlungsplanungen im Mittelpunkt, indem Ideen entwickelt werden müssen, die den Lernprozess der Kinder fördern (z. B. durch Impulse geben: „Wer von euch ist größer?“). Die Begrenzung auf Planungen wurde vor allem aus forschungspragmatischen Gründen getroffen, da sich die Zielgruppe der Untersuchung noch in der Ausbildung befindet, in der Praxisanteile sehr unterschiedlich integriert und nicht zwingend auf mathematische Themen bezogen sind.

2. Forschungsfragen

Mit der hier vorgestellten Studie wird die prognostische Validität des KomMa-Leistungstests zur Erfassung mathematikdidaktischen Wissens angehender Erzieher/innen geprüft. Um zu erfassen, ob das Ansinnen dieses Tests, eine für zukünftiges Handeln relevante Facette professionellen Wissens empirisch zu erfassen, tatsächlich zutrifft, wird als Prüfkriterium die Fähigkeit zur Handlungsplanung der angehenden Erzieher/innen videobasiert erhoben. In Bezug auf den Zusammenhang von Wissen und Handlungsplanung wird unter Bezug auf die dargestellten frühpädagogischen Konzeptualisierungen dabei angenommen, dass die Fähigkeit zur Situationswahrnehmung eine partiell oder vollständig vermittelnde Funktion übernimmt. Auf der Basis des Forschungsstandes zu Lehrkräften wird darüber hinaus ein direkter Effekt des mathematischen Wissens auf die Situationswahrnehmung angenommen. Abbildung 2 stellt ein entsprechendes Strukturmodell dar.

In der vorliegenden Studie werden konkret zwei konkurrierende Modelle geprüft und auf ihre Passung zu den Daten verglichen. Modell 1 nimmt einen direkten Effekt

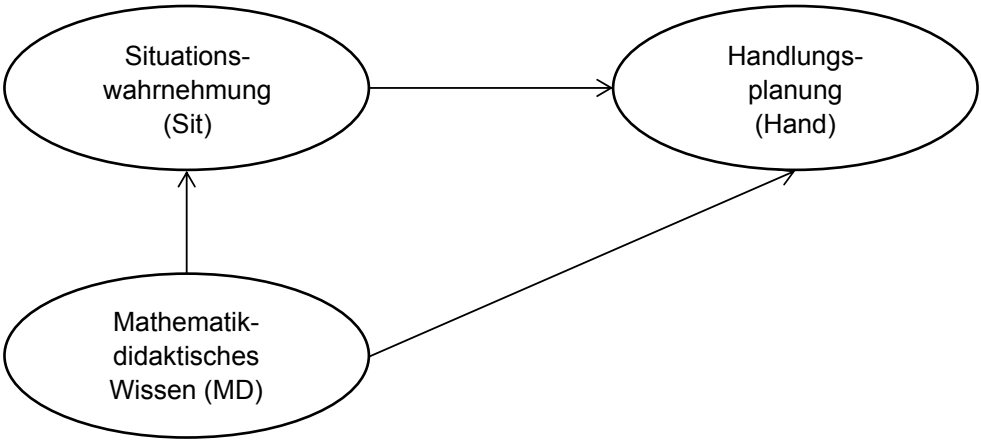


Abb. 2: Strukturmodell zum Verhältnis von mathematikdidaktischem Wissen (MD), der Fähigkeit zur Situationswahrnehmung (Sit) und der Fähigkeit zur Handlungsplanung (Hand)

des mathematikdidaktischen Wissens (MD) auf die Handlungsplanung (Hand) und die Situationswahrnehmung (Sit) sowie einen indirekten Effekt von MD auf Hand, vermittelt über Sit, an. Dieses Modell wurde direkt aus dem theoretischen Kompetenzmodell von Fröhlich-Gildhoff et al. (2011, S. 17–18) abgeleitet. Modell 2 prüft ein konkurrierendes Modell, indem lediglich ein indirekter Einfluss des mathematikdidaktischen Wissens, vermittelt über die Situationswahrnehmung, auf die Handlungsplanung angenommen wird (Thonhauser, 2007). In Modell 2 gilt folglich, dass der direkte Pfad von MD zu Hand in Abbildung 2 gleich null ist.

3. Methode

3.1 Instrumente

Leistungstest zur Erfassung des mathematikdidaktischen Wissens

Der KomMa-Leistungstest zur Erfassung des mathematikdidaktischen Wissens beruht auf einem Kompetenzmodell, das anhand einer Analyse der Bildungspläne und einschlägiger Literatur entwickelt und weiter oben beschrieben wurde. Die Itemkonstruktion folgte ebenfalls diesen Elementen (Jenßen et al., im Druck). Da die vorliegende Stichprobe im Verhältnis zur Anzahl der am Ende zu schätzenden Parameter relativ klein ist, wurde für diesen Beitrag ein Kurztest herangezogen, der aus zwölf Items besteht. Diese sind überwiegend im Multiple-Choice- und teilweise im offenen Format gestellt. Alle Items wurden für die Auswertungen dichotom (richtig/falsch) codiert. Für offene Items bildeten Codier-Anweisungen die Grundlage, die während der Itemkonstruktion entwickelt und im Rahmen von Prä-Tests und Pilotierungen überarbeitet worden waren. Die Interraterreliabilität erreicht mit Cohens $\kappa = 0.88$ für ein komplexes Kon-

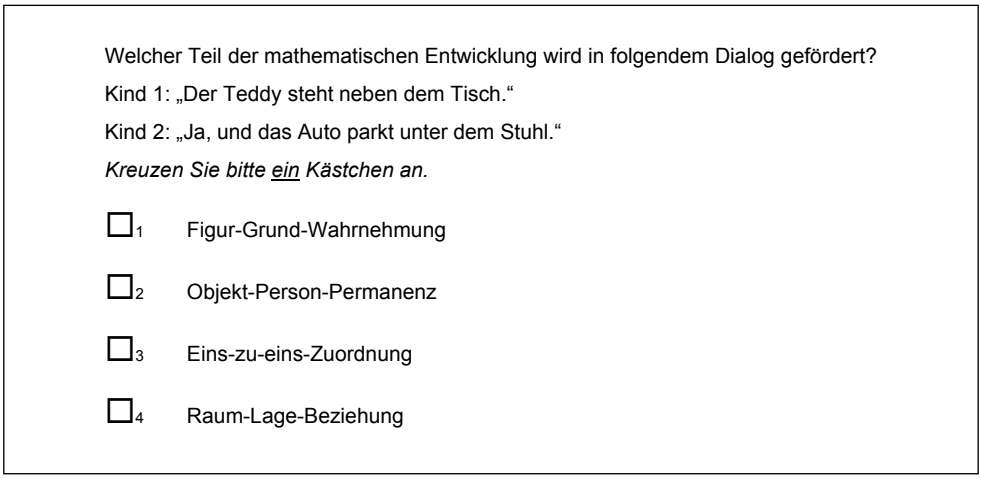


Abb. 3: Beispielitem aus dem KomMa-Leistungstest (richtige Antwort: Raum-Lage-Beziehung)

strukt wie mathematikdidaktisches Wissen einen guten Wert (Wirtz & Caspar, 2002, S. 59). Abbildung 3 zeigt ein Beispielitem im Multiple-Choice-Format.

Videobasierter Test zur Erfassung von Situationswahrnehmung und Handlungsplanung

Das Außenkriterium Handlungsplanung und der potenzielle Mediator Situationswahrnehmung wurden über einen videogestützten Test erfasst. Videos werden vor allem in der schulischen Unterrichtsforschung bereits seit Längerem eingesetzt, da sie eine standardisierte Datenerfassung ermöglichen und gleichzeitig eine hohe Situationsspezifität erreichen (König & Lebens, 2012; Baer et al., 2007). Darüber hinaus ist die Akzeptanz bei den Teilnehmer/innen höher als ein „klassischer" Papier-und-Bleistift-basierter Leistungstest (Seidel & Prenzel, 2007).

Aufgrund des Alters der Kinder wird in dieser Studie mit Videosequenzen gearbeitet, die Realsituationen aus einer Kindertagesstätte zeigen. Ansonsten folgt der Aufbau des Instruments dem, wie er auch in der Unterrichtsforschung gebräuchlich ist (König et al., 2014; Blomberg et al., 2011). Es wird mit drei Videos gearbeitet, die unterschiedliche mathematische Inhalte („Zahldarstellungen", „Messen" und „Bauen und Konstruieren") und Kontexte aus dem Alltag von Kindertagesstätten (Interaktion Erzieherin/Kind bzw. Freispiel) zeigen. Die Videos dienen als Ausgangspunkt, an den sich mehrere Items anschließen, die bei allen Videos identisch sind. Zunächst werden Multiple-Choice- und offene Items zur mathematikbezogenen Situationswahrnehmung bearbeitet. Dann wird allen angehenden Erzieher/innen erneut eine spezifische Situation aus dem Video gezeigt, und es werden offene Items zur Handlungsplanung bearbeitet.

Insgesamt besteht der Fragebogen aus zwölf Items zur Situationswahrnehmung (vier pro Video; z. B. „Nennen Sie drei mathematikdidaktische Aspekte der Situation und belegen Sie sie mit Beispielen") und zwölf Items zur Handlungsplanung (vier pro Vi-

deo; z. B. „Nennen Sie zwei mathematikdidaktisch angemessene Möglichkeiten, in dieser Situation zu reagieren“). Die Befragung einer Expertengruppe hat gezeigt, dass der Fragebogen als inhaltlich valide angesehen werden kann. Wie im KomMa-Leistungstest wurden alle Items dichotom (richtig/falsch) codiert. Die Codier-Anweisungen wurden anhand der Lösungen der Expert/innen, der Antworten aus einem Prä-Test (n = 15) und der einschlägigen Fachliteratur entwickelt. Im letzten Schritt wurden diese während der Hauptstudie um Antworten der Teilnehmer/innen ergänzt, sofern diese anhand von Fachliteratur validiert werden konnten. Die Interraterreliabilität erreichte mit Yules $Y \geq 0.8$ sehr gute Werte.[2]

3.2 Datenerhebung

Die Datenerhebung fand an Fachschulen für Sozialpädagogik in Niedersachsen und Berlin statt. Insgesamt wurden 16 Klassen getestet. Die angehenden Erzieher/innen bearbeiteten an zwei unterschiedlichen Messzeitpunkten zunächst den KomMa-Leistungstest zum mathematikdidaktischen Wissen und dann den videobasierten Fragebogen. Zwischen den beiden Messzeitpunkten lagen einige Tagen bis zwei Wochen, je nach den organisatorischen Möglichkeiten der Schulen. An beiden Messzeitpunkten wurden weitere Instrumente eingesetzt. Die Erhebungen wurden durch die Autor/innen bzw. geschulte Mitarbeiter/innen des Projekts durchgeführt, sodass die Durchführungsobjektivität gewährleistet ist.

3.3 Stichprobe

Die Stichprobe besteht aus 354 angehenden Erzieher/innen. Im Durchschnitt waren diese 23 Jahre alt (SD = 4.11; min = 17, max = 46). 83 % der Teilnehmer/innen waren weiblich und 17 % männlich, was den Verhältnissen in der Population in vollem Umfang entspricht (BMFSFJ, 2010, S. 13). Da es sich um eine Gelegenheitsstichprobe handelt und das komplexe Forschungsdesign mit mehreren Messzeitpunkten eine hohe Belastung für die teilnehmenden Schulen darstellte, sind die angehenden Erzieher/innen in unterschiedlichen Ausbildungsjahren. 41.5 % befinden sich im ersten Ausbildungsjahr, 33 % im zweiten und 25.5 % im dritten bzw. vierten Ausbildungsjahr.[3]

2 Da bei einigen Items ungleiche Randverteilungen vorlagen, wird Yules Y als Schätzung für Cohens κ berichtet (Wirtz & Caspar, 2002, S. 105).

3 Die Länge der Ausbildung unterscheidet sich in den beiden Bundesländern (Metzinger, 2006).

3.4 Datenanalyse

Die Datenanalysen wurden mit Mplus 5.2 (Muthén & Muthén, 2007) durchgeführt, wobei berücksichtigt wurde, dass die Teilnehmer/innen in Ausbildungsklassen getestet wurden, also eine geclusterte Datenstruktur vorliegt. Nach einer Betrachtung der Rohdaten wurden zunächst konfirmatorische Faktorenanalysen für die drei latenten Konstrukte mathematikdidaktisches Wissen, Situationswahrnehmung und Handlungsplanung geschätzt. Für jedes Modell wurde anhand der Passung an die Daten geprüft, ob Ladungen und/oder Intercepts fixiert werden können, um das beste Modell zu bestimmen. Im zweiten Schritt wurden die am besten passenden Modelle zu dem in Abbildung 2 dargestellten Strukturgleichungsmodell zusammengeführt. Da die Modelle nicht jeweils exakt dieselben Eigenschaften aufweisen (z. B. alle Ladungen fixiert und Intercepts frei geschätzt), resultiert ein Gesamtmodell, welches aus unterschiedlichen Teilmodellen besteht. Dies gilt es bei der Interpretation der Daten zu berücksichtigen. Die oben beschriebenen konkurrierenden Gesamtmodelle – direkter und indirekter Effekt von MD auf Hand vs. lediglich indirekter Effekt von MD auf Hand – wurden mithilfe eines χ^2-Differenztests auf ihre Passung an die Daten hin verglichen.

4. Ergebnisse

4.1 Deskriptive Ergebnisse

Die Summenwerte für die drei latenten Konstrukte sind in Tabelle 1 dargestellt. Alle drei Skalen können theoretisch Werte von 0 bis 12 annehmen.

Die angehenden Erzieher/innen erreichen beim mathematikdidaktischen Wissen und der Situationswahrnehmung im Mittel etwas höhere Werte als bei der Handlungsplanung. Minima und Maxima deuten in allen drei Fällen auf breite Spannweiten der Ergebnisse hin, da es Teilnehmer/innen gibt, die fast alle Items einer Skala richtig lösen können, und andere, die kaum ein Item richtig lösen.

Konstrukt	min	max	M	SD
Mathematikdidaktisches Wissen (MD)	0	11	6.8	2.40
Situationswahrnehmung (Sit)	1	11	6.3	2.01
Handlungsplanung (Hand)	0	11	4.3	2.65

Anmerkungen. min = Minimum, max = Maximum, M = Mittelwert, SD = Standardabweichung.

Tab. 1: Summenwerte der drei untersuchten Konstrukte

Mathematikdidaktisches Wissen

Im ersten Schritt wurde das am besten zu den Daten passende Modell für das mathematikdidaktische Wissen identifiziert. Bei zehn von zwölf Items konnten die Ladungen nicht als gleichwertig fixiert werden. Dieses Ergebnis deutet darauf hin, dass die Items das Konstrukt unterschiedlich stark repräsentieren. Der beste Modellfit wurde erreicht, wenn auch die Intercepts frei geschätzt wurden ($\chi^2(56) = 52.27$, $p = 0.62$, RMSEA = 0.00 [0.00; 0.03], SRMR = 0.04, CFI = 1.00). Dieses Modell passte auch signifikant besser als jenes, bei dem nur die Ladungen frei geschätzt wurden ($\Delta\chi^2(2) = 0.03$, $p = 0.99$). Die Modellwerte sind in Tabelle 2 dargestellt. Cronbach's α liegt bei akzeptablen 0.65. Die Varianz der Skala ist mit 0.02 ($p < 0.05$) zwar gering, aber signifikant.

Situationswahrnehmung

Für die Skala zur Situationswahrnehmung wurde ein Modell geschätzt, bei dem ein Summenwert über die drei eingesetzten Videos gebildet wurde. Dieses sog. Item-Parceling wird vor allem dann empfohlen, wenn primär die Struktur der latenten Konstrukte von Interesse ist (Little, Cunningham, Shahar & Widaman, 2002) und wenn es sich um eine relativ kleine Stichprobe im Verhältnis zur Anzahl der zu schätzenden Parameter handelt (Bandalos & Finney, 2001, S. 270). Beides trifft auf die hier vorliegende Studie zu. Die Faktorladungen und Intercepts konnten fixiert werden ($\chi^2(3) = 3.07$, $p = 0.38$, RMSEA = 0.01 [0.00; 0.11], SRMR = 0.04, CFI = 0.99). Auch ein Modell, in dem nur

Indikator	λ	SE(λ)	p-Wert	R^2	SE(R^2)	p-Wert
Zahlzerlegung der 10	0.30	0.05	0.000	0.09	0.03	0.003
Dreiecksbegriff	0.47	0.05	0.000	0.22	0.05	0.000
mathematische Raumgestaltung	0.26	0.05	0.000	0.07	0.02	0.005
Beobachtung	0.40	0.05	0.000	0.16	0.04	0.000
Zahlen im Würfelspiel	0.38	0.05	0.000	0.15	0.04	0.000
Erfahrungen zum Messen	0.37	0.07	0.000	0.14	0.05	0.008
Mustersinn	0.26	0.05	0.000	0.07	0.02	0.005
aktiv-entdeckendes Lernen in der Förderung	0.40	0.04	0.000	0.16	0.04	0.000
Raum-Lage-Beziehungen	0.49	0.08	0.000	0.24	0.08	0.004
geometrisches Denken	0.30	0.07	0.000	0.09	0.04	0.023
mathematische Angebote (geplant oder alltagsintegriert)	0.38	0.05	0.000	0.15	0.04	0.000
Sortieren und Klassifizieren	0.35	0.07	0.000	0.12	0.05	0.007

Anmerkungen. λ = Faktorladung, SE(λ) = Standardfehler der Faktorladung, R^2 = Varianzaufklärung.

Tab. 2: Faktorladungen und Varianzaufklärung des Modells für mathematikdidaktisches Wissen (standardisierte Werte aus einer konfirmatorischen Faktorenanalyse)

Indikator	λ	SE(λ)	p-Wert	R^2	SE(R^2)	p-Wert
Zahldarstellungen (Video 1)	0.56	0.05	0.000	0.31	0.06	0.000
Messen (Video 2)	0.47	0.04	0.000	0.22	0.03	0.000
Bauen und Konstruieren (Video 3)	0.52	0.04	0.000	0.27	0.04	0.000

Anmerkungen. λ = Faktorladung, SE(λ) = Standardfehler der Faktorladung, R^2 = Varianzaufklärung.

Tab. 3: Faktorladungen und Varianzaufklärung im Modell für Situationswahrnehmung (standardisierte Werte aus einer konfirmatorischen Faktorenanalyse)

die Faktorladungen fixiert wurden, erreichte akzeptable Modellfit-Werte ($\chi^2(2) = 3.22$, $p = 0.20$, RMSEA = 0.05 [0.00; 0.14], SRMR = 0.04, CFI = 0.99). Der Modellvergleich mit dem χ^2-Differenztest bestätigte allerdings, dass das erste Modell besser zu den vorliegenden Daten passt ($\Delta\chi^2(1) = 0.46$, $p = 0.50$). Die Modellwerte des identifizierten Modells sind in Tabelle 3 dargestellt. Dass das Fixieren der Ladungen zu einem besseren Gesamtfit führt, deutet darauf hin, dass die Fähigkeit zur Situationswahrnehmung über alle mathematischen Bildungsinhalte hinweg ähnlich gelingt. Für diese Skala hat sich ein relativ niedriges Cronbach's α von 0.53 ergeben. Die Varianz der Skala ist mit 0.23 (0.03, $p < 0.001$) ebenfalls niedrig, aber signifikant.

Handlungsplanung

Als drittes wurde ein Modell für die Handlungsplanung geschätzt. Auch hier dienen die drei eingesetzten Videos als Indikatoren. Der beste Modellfit konnte erreicht werden, wenn die Faktorladungen frei geschätzt und die Intercepts fixiert wurden ($\chi^2(2) = 4.85$, $p = 0.09$, RMSEA = 0.07 [0.00; 0.16], SRMR = 0.04, CFI = 0.98). Das Modell, in dem zusätzlich die Faktorladungen fixiert wurden, erreichte keine akzeptablen Fit-Werte ($\chi^2(4) = 88.61$, $p = 0.00$, RMSEA = 0.28 [0.23; 0.34], SRMR = 0.16, CFI = 0.37). Dieses Ergebnis deutet darauf hin, dass die eingesetzten Videos das Konstrukt unterschiedlich stark repräsentieren. Die Einzelwerte dieses Modells sind in Tabelle 4 dargestellt. Cronbach's α liegt für diese Skala bei guten 0.70. Auch in diesem Modell ist die Varianz mit 0.27 (0.05, $p < 0.001$) niedrig, aber signifikant.

Indikator	λ	SE(λ)	p-Wert	R^2	SE(R^2)	p-Wert
Zahldarstellungen (Video 1)	0.51	0.04	0.000	0.26	0.04	0.000
Messen (Video 2)	0.73	0.04	0.000	0.54	0.05	0.000
Bauen und Konstruieren (Video 3)	0.70	0.03	0.000	0.48	0.04	0.000

Anmerkungen. λ = Faktorladung, SE(λ) = Standardfehler der Faktorladung, R^2 = Varianzaufklärung.

Tab. 4: Faktorladungen und Varianzaufklärung im Modell für Handlungsplanung (standardisierte Werte aus einer Faktorenanalyse)

Zusammenhänge zwischen Wissen und Handlungsplanung

Die am besten passenden Modelle der drei Konstrukte wurden schließlich in einem linearen Strukturgleichungsmodell zusammengeführt und auf ihre Zusammenhänge hin geprüft. In Modell 1 wird sowohl ein direkter als auch ein indirekter Effekt des mathematikdidaktischen Wissens auf die Handlungsplanung angenommen. Im konkurrierenden Modell 2 wird nur der indirekte Effekt des mathematikdidaktischen Wissens, vermittelt über die Situationswahrnehmung, angenommen. Dies entspricht dem Produkt der Pfadkoeffizienten β_{21} und β_{23} in Abbildung 4 und der Fixierung von β_{31} auf null.

Die Analyse zeigt, dass der in Modell 1 angenommene direkte Effekt des mathematikdidaktischen Wissens auf Hand nicht signifikant ist. Im Weiteren wird deswegen nur das Modell 2 weiter verfolgt, welches einen indirekten Effekt postuliert. Der direkte Effekt des mathematikdidaktischen Wissens auf die Handlungsplanung wurde in diesem Modell auf null fixiert, womit es sparsamer ist als Modell 1. Die Passung dieses Modells an die Daten ist zufriedenstellend ($\chi^2(140) = 169.27$, $p = 0.05$, RMSEA = 0.03 [0.00; 0.04], SRMR = 0.06, CFI = 0.95) (Schermelleh-Engel, Moosbrugger & Müller, 2003).

Abbildung 4 zeigt die Regressionsgewichte in dem Modell. Es zeigen sich signifikante und starke direkte Effekte des mathematikdidaktischen Wissens auf die Situationswahrnehmung sowie sehr starke Effekte von der Situationswahrnehmung auf die Handlungsplanung, die sich kaum voneinander trennen lassen. Dies könnte methodisch

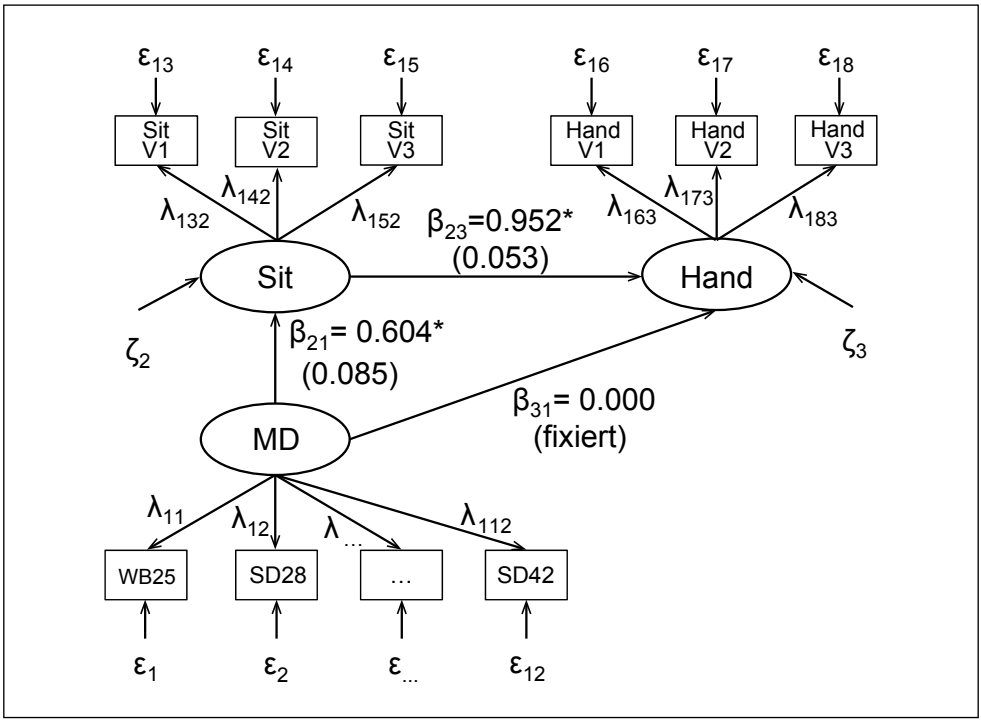


Abb. 4: Lineares Strukturgleichungsmodell mit Regressionsgewichten

darauf hindeuten, dass es sich um ein Konstrukt handelt. Aufgrund der Konzeptualisierung werden diese aber weiterhin als getrennte Konstrukte behandelt.

Der indirekte Effekt von MD auf Hand beträgt $\beta_{ind} = 0.58$ (SE = 0.07, $p < 0.001$) und kann als hoch bezeichnet werden. Eine Signifikanzprüfung mithilfe der Bootstrapping-Methode (MacKinnon, 2008) zeigt über das Konfidenzintervall [0.41; 0.74], dass der indirekte Effekt systematisch von 0 verschieden ist ($p < 0.001$).[4]

5. Diskussion und Ausblick

Im Kontext der Lehrerausbildungsforschung hat sich gezeigt, dass das mathematische, mathematikdidaktische und allgemeinpädagogische Wissen von zentraler Bedeutung für die Situationswahrnehmung und Handlungsplanung von Lehrkräften ist (Blömeke, König et al., 2014). Für Erzieher/innen fehlen bislang entsprechende Hinweise (Thole, 2010). Zur Schließung dieser Forschungslücke wurden in der hier präsentierten Studie zwei Instrumente, ein Papier-und-Bleistift-basierter Leistungstest zur Erfassung des mathematikdidaktischen Wissens und ein videobasierter Fragebogen zur Erfassung der Situationswahrnehmung und Handlungsplanung von angehenden Erzieher/innen, eingesetzt.

Beiden Instrumenten konnte im Zuge des Konstruktionsprozesses durch Expertenbefragungen inhaltliche Validität bescheinigt werden. Mit einem linearen Strukturgleichungsmodell zeigen wir nun in diesem Beitrag, dass das mathematikdidaktische Wissen direkt die Situationswahrnehmung und indirekt auch die Handlungsplanung vorhersagt. Ein direkter Effekt des mathematikdidaktischen Wissens auf die Handlungsplanung kann dagegen nicht nachgewiesen werden, sondern dieser Einfluss wird über die Situationswahrnehmung vermittelt. Diese Ergebnisse stehen in Einklang mit Erkenntnissen aus der Lehrerbildung (Thonhauser, 2007). Sie deuten darauf hin, dass es sich um zwei Prozesse handelt, die eher aufeinander folgen, als dass sie parallel stattfinden bzw. ineinander verwoben sind.

Wenn er im Einklang mit den Vorannahmen steht, kann ein indirekter Effekt ein ebenso starker Hinweis für kriteriale Validität sein wie ein direkter (Cronbach & Meehl, 1955). Da der Nachweis der prognostischen Validität als eine Facette kriterialer Validität eine zentrale Forschungsfrage dieser Studie ist, stellen unsere Ergebnisse somit einen starken Hinweis auf die Validität des KomMa-Leistungstests dar, da unabhängig von der konkreten Modellierung als direkter oder indirekter Effekt die Situationswahrnehmung und die Handlungsplanung der angehenden Erzieher/innen signifikant und mit hohen Effektstärken vorhergesagt werden.

4 Da der Bootstrap-Befehl in MPlus nicht in Kombination mit dem Cluster-Befehl verwendet werden kann (Muthén & Muthén, 2007, S. 496), wurde dieser für diese Analyse entfernt. Dies ist insofern vertretbar, da der Scaling Correction Factor mit 0.98 nahe 1 liegt, d. h. die Klassen unterscheiden sich nicht wesentlich voneinander.

In Bezug auf die hier präsentierten Ergebnisse müssen allerdings auch einige Grenzen diskutiert werden. So ist zu berücksichtigen, dass ein komplexes Konstrukt wie mathematikdidaktisches Wissen mit 12 Items möglicherweise nicht voll erfasst ist. Dieser Kurztest wurde für den vorliegenden Beitrag wegen der vergleichsweise kleinen Stichprobe verwendet. Wir nehmen derzeit zusätzliche Erhebungen vor, sodass der Test in seiner Langform (30 Items) für zukünftige Skalierungen verwendet werden kann.

Zu den videobasierten Erhebungen ist als Qualität festzuhalten, dass unterschiedliche mathematische Bildungsinhalte und Situationen aus dem Kita-Alltag gezeigt werden und die Testung so situationsbezogen und realistisch ist. Allerdings ist nicht geklärt, inwieweit von den im Instrument eingesetzten Situationen auf andere Situationen generalisiert werden kann (Kane, 1992). Besonders zu berücksichtigen ist zudem das hohe Regressionsgewicht zwischen der Fähigkeit zur Situationswahrnehmung und der Fähigkeit zur Handlungsplanung, das diese empirisch gesehen als kaum unterscheidbar erscheinen lässt. Dieses Ergebnis schränkt die oben getroffene Aussage ein, dass es sich um sequenzielle Prozesse handele, könnte aber möglicherweise darauf zurückzuführen sein, dass beide Merkmale aus forschungsökonomischen Gründen mit derselben Methode – also demselben Fragebogen, identischen Videos und zum selben Messzeitpunkt – erhoben wurden. Zukünftig sollten die Konstrukte besser mit unterschiedlichen Videos erfasst werden.

In Bezug auf das Forschungsdesign (Cronbach & Meehl, 1955) ist angesichts von drei Messzeitpunkten auf mögliche Motivationsprobleme der Teilnehmer/innen hinzuweisen, die allerdings nur relevant wären, wenn es bei einer Teilpopulation zu systematischen Verzerrungen bei der Testbearbeitung gekommen sein sollte (z. B. bewusstes Überspringen von Aufgaben). Und schließlich ist darauf hinzuweisen, dass die hier untersuchte Population kaum spezifische Lerngelegenheiten im Bereich Mathematik und ihrer Didaktik hatte, die aber für die Bearbeitung des Tests erforderlich sind (Dunekacke et al., 2014). Entsprechend ist eine Folge, dass alle Modelle unserer Studie nur eine geringe Varianz und die Skala zur Situationswahrnehmung eine relativ niedrige Reliabilität aufweisen, da die Stichprobe im Hinblick auf MD, Sit und Hand sehr homogen ist (Bühner, 2011). Ob dies auf eine Selektion bei der Ziehung der hier berichteten Gelegenheitsstichprobe zurückzuführen ist oder ob es sich um eine generelle Einschränkung in der Population aufgrund tatsächlich geringer Lerngelegenheiten handelt, muss in weiteren Studien überprüft werden.

Diese Diskussion weiterführend fällt auf, dass die Faktorladungen beim mathematikdidaktischen Wissen und der Fähigkeit zur Situationswahrnehmung teilweise niedrig sind. Dieses könnte ein Hinweis darauf sein, dass es sich um eher heterogene Konstrukte handelt. Hierfür spricht auch, dass die Korrelationen unter den Items teilweise niedrig sind und eine hohe Residualvarianz bzw. niedrige Determination mit sich bringen. Auffällig ist außerdem, dass bei der Handlungsplanung das Video zur Zahldarstellung (Video 1) eine geringere Ladung als die beiden anderen Videos aufweist. Dies könnte darauf hindeuten, dass die Handlungsplanung im Bereich der Zahldarstellung andere Fähigkeiten als in den anderen beiden Bereichen erfordert bzw. dass sich die beiden anderen Bereiche stärker ähneln. Dies wiederum könnte möglicherweise daran

liegen, dass in den Bereichen Messen und Bauen und Konstruieren stärker handlungsorientiert und materialbasiert gearbeitet werden kann.

Unsere Studie gibt insofern auch interessante Hinweise auf die Natur der untersuchten Konstrukte bei Erzieher/innen. Da sich andeutet, dass auch für diese Population das professionelle Wissen ein zentraler Prädiktor professionellen Handelns ist, der durch die Wahrnehmung der Situation vermittelt wird, kann für den Lernbereich Mathematik in aller Vorsicht geschlussfolgert werden, dass es bedeutsam ist, entsprechende Lerngelegenheiten in der Ausbildung zur Verfügung zu stellen. Zugleich ist zu bedenken, dass theoretisch fundiertes fachdidaktisches Wissen offensichtlich vor allem dann sinnvoll angewendet werden kann, wenn die Spezifika der vorliegenden informellen Situation in einer Kindertageseinrichtung erkannt und genutzt werden können (von Balluseck & Nentwig-Gesemann, 2008). Damit ergibt sich, dass in der Ausbildung *„die Verschränkung von theoretischem Wissen und praktischen Handlungsvollzügen unabdingbar"* (ebd., S. 30) ist. Bislang wird der Lernbereich Mathematik im Rahmen der Ausbildung sehr unterschiedlich thematisiert, wie eine im Projekt durchgeführte Analyse der Ausbildungsordnungen gezeigt hat (Jenßen et al., im Druck). Zukünftige Konzeptionen sollten daher in Ergänzung zu einer generellen Stärkung mathematikbezogener Lerngelegenheiten die praxisbezogene Verschränkung von fach- und fachdidaktischen Inhalten mit handlungspraktischen Erfahrungen bedenken, indem beispielsweise mathematische Probleme aus dem Kontext der Kita (z. B. das Bilden von Mustern) mehrperspektivisch betrachtet werden.

Literatur

Anders, Y. (2012). *Modelle professioneller Kompetenzen für frühpädagogische Fachkräfte. Aktueller Stand und ihr Bezug zur Professionalisierung. Expertise zum Gutachten „Professionalisierung in der Frühpädagogik"*. München: Aktionsrat Bildung.

Baer, M., Dörr, G., Fraefel, U., Kocher, M., Küster, O., Larcher, S., Müller, P., Sempert, W., & Wyss, C. (2007). Werden angehende Lehrpersonen durch das Studium kompetenter? – Kompetenzaufbau und Standarderreichung in der berufswissenschaftlichen Ausbildung an drei Pädagogischen Hochschulen in der Schweiz und in Deutschland. *Unterrichtswissenschaft, 35*(1), 15–47.

Ball, D., & Bass, H. (2009). With an Eye on the Mathematical Horizon: Knowing Mathematics for Teaching to Learners' Mathematical Futures. In M. Neubrand (Hrsg.), *Beiträge zum Mathematikunterricht 2009* (S. 11–22). Münster: WTM.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Hrsg.), *Advanced structural equation modeling: New developments and techniques* (S. 269–296). Mahwah: Erlbaum.

Blomberg, G., Stürmer, K., & Seidel, T. (2011). How pre-service teachers observe teaching on video: Effects of viewers' teaching subjects and the subject of the video. *Teaching and Teacher Education, 27*, 1131–1140.

Blömeke, S., Benthien, J., Döhrmann, M., Busse, A., Kaiser, G., & König, J. (im Druck). Teacher change during induction: Profiles in the development of beginning primary teachers' knowledge and beliefs and their relation to performance. *International Journal of Science and Mathematics Education*.

Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (re-submitted after revisions). On the Nature of Teacher Expertise: Modeling the Relations Between Knowledge, Perceptual Accuracy, and Speed. *American Educational Research Journal.*

Blömeke, S., Felbrich, A., Müller, C., Kaiser, G., & Lehmann, R. (2008). Effectiveness of teacher education. State of research, measurement issues and consequences for future studies. *ZDM – The International Journal on Mathematics Education, 40*(5), 719–734.

Blömeke, S., Kaiser, G., Döhrmann, M., Suhl, U., & Lehmann, R. (2010). 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. 195–252). Münster: Waxmann.

Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M., & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf: Fachbezogenes Wissen als Voraussetzung einer genauen Wahrnehmung und Analyse von Unterricht. *Zeitschrift für Erziehungswissenschaft, 17*(3), 509–542.

BMFSFJ Bundesministerium für Familie, Senioren, Frauen und Jugend (2010). *Männliche Fachkräfte in Kindertagesstätten – eine Studie zur Situation von Männern in Kindertagestätten und in der Ausbildung zum Erzieher*. Berlin: Sinus Sociovision.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Aufl.). Heidelberg: Springer.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktual. Aufl.). München: Pearson.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School Readiness and Later Achievement. *Developmental psychology, 43*(6), 1428–1462.

Dunekacke, S., Buhl, M., Jenßen, L., Baack, W., Grassmann, M., & Blömeke, S. (2014). Mathematisches Fachwissen von angehenden Erzieher/-innen und Grundschullehrer/-innen im Vergleich. *Symposium – Perspektiven mathematischer Bildung im Übergang vom Kindergarten zur Grundschule*. Freiburg.

Fried, L., & Roux, S. (2009). Zur Pädagogik der frühen Kindheit im 21. Jahrhundert – Desiderata. In L. Fried & S. Roux (Hrsg.), *Pädagogik der frühen Kindheit. Handbuch und Nachschlagewerk* (S. 378–382). Berlin: Cornelsen.

Fröhlich-Gildhoff, K., Nentwig-Gesemann, I., & Pietsch, S. (2011). *Kompetenzorientierung in der Qualifizierung frühpädagogischer Fachkräfte. Eine Expertise der Weiterbildungsinitiative Frühpädagogische Fachkräfte (WiFF).* München: Deutsches Jugendinstitut e. V.

Gasteiger, H. (2010). *Elementare mathematische Bildung im Alltag der Kindertagesstätte. Grundlegung und Evaluation eines kompetenzorientierten Förderansatzes*. Münster: Waxmann.

Ginsburg, H. P., & Ertle, B. (2008). Knowing the Mathematics in Early Childhood Mathematics. In O. N. Saracho & B. Spodek (Hrsg.), *Contemporary Perspectives on Mathematics in Early Childhood Education* (S. 45–66). Charlotte: Information AGE.

Gudjons, H. (2008). *Handlungsorientiert lehren und lernen: Schüleraktivierung – Selbsttätigkeit – Projektarbeit* (7., aktual. Aufl.). Bad Heilbrunn: Klinkhardt.

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 143–171). Heidelberg: Springer.

Hogrebe, N., Schulz, S., & Böttcher, W. (2012). Professionalisierung im Elementarbereich – Personalentwicklung im Spannungsfeld von Anspruch und Wirklichkeit. *Soziale Passagen, 4*, 247–261.

Jenßen, L., Dunekacke, S., Baack, W., Tengler, M., Koinzer, T., Schmude, C., Wedekind, H., Grassmann, M., & Blömeke, S. (im Druck). *KomMa: Mathematikbezogene Kompetenz von*

Erzieher/-innen: Theoretischer Rahmen, Strukturanalyse und Zusammenhang zu Ausbildungsinhalten.

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik, 61. Beiheft*, 11–31.

Joas, H. (1996). *Die Kreativität des Handelns*. Frankfurt a. M.: Suhrkamp.

Kane, M. T. (1992). The assessment of professional competence. *Evaluation & the Health Professions, 15*(2), 163–182.

Klibanoff, R. S., Levine, S. C., Huttenlocher, J., Vasilyeva, M., & Hedges, L. V. (2006). Preschool Children's Mathematical Knowledge: The Effect of Teacher „Math Talk". *Developmental Psychology, 42*(1), 56–69.

König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A., & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education, 38*, 76–88.

König, J., & Lebens, M. (2012). Classroom Management Expertise (CME) von Lehrkräften messen: Überlegungen zur Testung mithilfe von Videovignetten und erste empirische Befunde. *Lehrerbildung auf dem Prüfstand, 5*(1), 3–29.

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction, 19*, 513–526.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151–173.

Lee, J. (2010). Exploring Kindergarten Teachers' Pedagogical Content Knowledge of Mathematics. *International Journal of Early Childhood, 47*(1), 27–41.

Lee, J., Meadows, M., & Lee, J. O. (2003). *What causes teachers to implement high quality mathematics education more frequently: Focusing on teachers' pedagogical content knowledge*. Washington, D. C.: ERIC.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah: Erlbaum.

Metzinger, A. (2006). Geschichte der Erzieherinnenausbildung als Frauenberuf. In L. Fried & S. Roux (Hrsg.), *Pädagogik der frühen Kindheit. Handbuch und Nachschlagewerk* (S. 348–358). Weinheim/Basel: Beltz.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide. Fifth Edition*. Los Angeles: Muthén & Muthén.

National Advisory Panel (2008). *The Final Report of the National Mathematics Advisory Panel*. U. S. Department of Education.

Perrez, M., Huber, G. L., & Geißler, K. A. (2001). Psychologie der pädagogischen Interaktion. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (4. vollst. überarb. Aufl., S. 358–413). Weinheim/Basel: Beltz.

Reynolds, A. (1995). One Year of Preschool Intervention or Two: Does it Matter? *Early Childhood Research Quarterly, 10*, 1–31.

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

Sarama, J., & Clements, D. H. (2009). *Early Childhood Mathematics Education Research. Learning Trajectories for Young Children*. New York: Routledge.

Schäfer, G. E. (2005). *Überlegungen zur Professionalisierung von Erzieherinnen.* http://www.bosch-stiftung.de/content/language1/downloads/rahmencurriculum_schaefer.pdf [24. 07. 2014].

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online, 8*(2), 23–74.

Seidel, T., & Prenzel, M. (2007). Wie Lehrpersonen Unterricht wahrnehmen und einschätzen – Erfassung pädagogisch-psychologischer Kompetenzen mit Videosequenzen. In M. Prenzel,

I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (Zeitschrift für Erziehungswissenschaft, Sonderheft 8, S. 201–216). Wiesbaden: VS Verlag für Sozialwissenschaften.

Sherin, M. G. (2007). The development of teachers' professional vision in video clubs. In R. Goldman, R. Pea, B. Barron & S. J. Derry (Hrsg.), *Video research in the learning sciences* (S. 383–395). Mahwah: Lawrence Erlbaum.

Shulman, L. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*(2), 4–14.

Star, J. R., & Strickland, S. K. (2008). Learning to observe: Using video to improve preservice mathematics teachers' ability to notice. *Journal for Mathematics Teacher Education, 11*(2), 107–125.

Thole, W. (2010). Die pädagogischen MitarbeiterInnen in Kindertageseinrichtungen. Professionalität und Professionalisierung eines pädagogischen Arbeitsfeldes. *Zeitschrift für Pädagogik, 53*(2), 206–222.

Thonhauser, J. (2007). Lehrer/-innen handeln situationsspezifisch. In A. Gastager, T. Hascher & H. Schwetz (Hrsg.), *Pädagogisches Handeln: Balance zwischen Theorie und Praxis. Beiträge zur Wirksamkeitsforschung in pädagogisch-psychologischem Kontext. Erziehungswissenschaft, Bd. 24* (S. 47–60). Landau: VEP.

van Es, E. A., & Sherin, M. G. (2006). Mathematics teachers' „learning to notice" in the context of a video club. *Teaching and Teacher Education, 24*, 244–276.

van Oers, B. (2009). Emergent mathematical thinking in the context of play. *Educational Studies in Mathematics, 74*(1), 23–37.

von Balluseck, H., & Nentwig-Gesemann, I. (2008). Wissen Können Reflexion. Die Verbindung von Theorie und Praxis in der Ausbildung von Erzieherinnen. *Sozial Extra, 32*(3-4), 28–32.

Weinert, F. E. (2001). Concept of Competence: A Conceptual Classification. In D. S. Rychen & L. Hersh Salganik (Hrsg.), *Defining and Selecting Key Competencies*. Göttingen: Hogrefe.

Weinert, F. E., Schrader, F.-W., & Helmke, A. (1990). Unterrichtsexpertise: Ein Konzept zur Verringerung der Kluft zwischen zwei theoretischen Paradigmen. In L.-M. Alisch, J. Baumert & K. Beck (Hrsg.), *Professionswissen und Professionalisierung: Sonderband in Zusammenarbeit mit der Zeitschrift Empirische Pädagogik* (Braunschweiger Studien zur Erziehungs- und Sozialarbeit, 28, S. 173–206). Braunschweig: Copy-Center Colmesee.

Widulle, W. (2009). *Handlungsorientiert Lernen im Studium. Arbeitsbuch für soziale und pädagogische Berufe*. Heidelberg: Springer.

Wild, K.-P., & Krapp, A. (2001). Pädagogisch-psychologische Diagnostik. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (4. vollst. überarb. Aufl., S. 513–563). Weinheim/Basel: Beltz.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.

Abstract: This paper examines the prognostic validity of a paper-and-pencil test to assess the mathematical pedagogical content knowledge (MPCK) of prospective pre-school teachers. The test was developed in the KomMa project. Teacher education research revealed that MPCK significantly predicts the perception and performance of young teachers. Corresponding hypotheses were developed with respect to pre-school teachers. The skill to perceive pre-school situations and the skill to plan actions were assessed via a video-based assessment. Both instruments were applied to a sample of 354 prospective pre-school teachers. Competing structural equation models revealed two major findings: First, MPCK has a significant direct influence on the perception of pre-school situations. Second, MPCK indirectly influences performance. These results provide evidence of the prognostic validity of the achievement test and they shed light on the relation of knowledge, perception and performance.

Keywords: Preschool Teacher, Mathematics Pedagogical Content Knowledge, Validation, Achievement Test, Video-Based Assessment

Anschrift der Autorinnen/des Autors

M.A. Simone Dunekacke, Humboldt-Universität zu Berlin,
Institut für Erziehungswissenschaften, Abteilung Systematische Didaktik
und Unterrichtsforschung, und Carl von Ossietzky Universität Oldenburg,
Institut für Pädagogik, Uhlhornsweg, 26111 Oldenburg, Deutschland
E-Mail: simone.dunekacke@uni-oldenburg.de

Dipl.-Psych. Lars Jenßen, Humboldt-Universität zu Berlin,
Institut für Erziehungswissenschaften, Abteilung Systematische Didaktik
und Unterrichtsforschung, Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: lars.jenssen@hu-berlin.de

Prof. Dr. Sigrid Blömeke, Centre for Educational Measurement at the University of Oslo
(CEMO), Leibniz-Institut für die Pädagogik der Mathematik und Naturwissenschaften Kiel,
Humboldt-Universität zu Berlin, Postboks 1072/Blindern, 0316 Oslo, Norwegen
E-Mail: sigribl@cemo.uio.no

Franziska Bouley/Stefanie Berger/Sabine Fritsch/Eveline Wuttke/
Jürgen Seifried/Kathleen Schnick-Vollmer/Bernhard Schmitz

Der Einfluss von universitären und außeruniversitären Lerngelegenheiten auf das Fachwissen und fachdidaktische Wissen von angehenden Lehrkräften an kaufmännisch-berufsbildenden Schulen

Zusammenfassung: Fachwissen und fachdidaktisches Wissen gelten als bedeutsame Prädiktoren für erfolgreiches Lehrerhandeln. Die universitäre Lehrerbildung sollte daher Lerngelegenheiten anbieten, die angehenden Lehrkräften den Aufbau entsprechender Wissensbestände ermöglichen. Im vorliegenden Beitrag wird am Beispiel der Wirtschaftspädagogik der Einfluss von außeruniversitären und universitären Lerngelegenheiten auf Fachwissen und fachdidaktisches Wissen im Rechnungswesen analysiert. Die Befunde verweisen auf einen positiven Einfluss außeruniversitärer Lerngelegenheiten (insbesondere der kaufmännischen Erstausbildung) auf Fachwissen und fachdidaktisches Wissen. Universitäre fachwissenschaftliche und fachdidaktische Lerngelegenheiten scheinen dagegen eine vergleichsweise geringe Rolle zu spielen.

Schlagworte: Fachwissen, fachdidaktisches Wissen, Lerngelegenheiten, Lehrerbildung, Wirtschaftspädagogik

1. Problemstellung

Während in den letzten Jahren vermehrt Forschungsbemühungen zum Kompetenzerwerb sowie zur Kompetenzentwicklung von Schülerinnen und Schülern zu verzeichnen sind (z. B. PISA, TIMSS, IGLU oder DESI, für die duale berufliche Ausbildung siehe die BMBF-Forschungsinitiative ASCOT, z. B. BMBF, 2012; Seeber et al., 2010), fehlt es an empirischen Befunden zur Kompetenzentwicklung im tertiären Bildungssektor (Blömeke, Zlatkin-Troitschanskaia, Kuhn & Fege, 2013). Die Lehrerbildung kann dabei – zumindest im Ansatz – als Ausnahme gelten. Neben einer Reihe von Einzeluntersuchungen (z. B. Döbrich, Klemm, Knauss & Lange, 2003; Oser & Oelkers, 2001; Schaefers, 2002) sind v. a. internationale Vergleichsstudien der Lehrerbildungssysteme von Interesse (Eurydice, 2003, 2004; OECD, 2005). Insbesondere von den Studien TEDS-M (Blömeke, Kaiser & Lehmann, 2010) und COACTIV (Krauss et al., 2008; Kunter et al., 2011) gehen Impulse für die Diskussion evidenzbasierter Lehrerbildung aus (Hascher, 2011). Dabei können speziell der Bereich der Mathematik sowie – mit Abstrichen – jener der Naturwissenschaften als gut beschrieben gelten (Neuweg, 2014). Für die Wirksamkeit der universitären Lehrerbildung im kaufmännisch-berufsbildenden

Bereich liegen dagegen bislang kaum empirisch fundierte Aussagen vor (Kuhn et al., 2014; Zlatkin-Troitschanskaia, Förster, Brückner, Hansen & Happ, 2013).

Diese Forschungslücke ist Ausgangspunkt des Projektes „Kompetenzmodellierung und Kompetenzmessung im wirtschaftspädagogischen Studium“ (KoMeWP).[1] Es zielt auf die Modellierung und Erfassung der professionellen Kompetenz angehender Lehrkräfte an kaufmännisch-berufsbildenden Schulen ab und nimmt die Domäne des Rechnungswesens in den Blick, und zwar aus folgendem Grund: Dem betrieblichen Rechnungswesen als „Kern des betrieblichen Informationssystems“ (Preiß, 2000, S. 7) wird in der kaufmännischen Ausbildung große Bedeutung beigemessen, da dieser Inhaltsbereich Lernende über die Auseinandersetzung mit der reinen (Buchungs-)Systematik hinaus dazu befähigen soll, unternehmerische Prozesse zu durchdringen und ökonomisches Denken zu entwickeln (Achtenhagen, 1996). Einschlägige empirische Studien zeigen jedoch, dass angehende Lehrkräfte an kaufmännischen Schulen (Bachelor- und Masterstudierende der Wirtschaftspädagogik sowie Referendarinnen und Referendare) hinsichtlich der für die Gestaltung von Unterrichtsprozessen in dieser Domäne zentralen fachwissenschaftlichen und fachdidaktischen Kompetenzen Defizite aufweisen (vgl. z. B. Seifried, Türling & Wuttke, 2010; Türling, 2014; Wuttke & Seifried, 2013). Dies legt nahe, dass es zumindest in der ersten Phase der Lehrerbildung nicht in hinreichendem Maße gelingt, entsprechende Kompetenzen zu entwickeln (ähnliche Befunde werden ebenfalls für Mathematik und Naturwissenschaften berichtet, z. B. Abell, 2007; Ball, Lubienski & Mewborn, 2001; Halim & Meerah, 2002; Thanheiser, 2009). Möglicherweise fehlt es hier an entsprechenden Lerngelegenheiten im wirtschaftspädagogischen Studium. Für diese Annahme spricht eine jüngst durchgeführte, auf Durchsicht von Modulhandbüchern, Studien- und Prüfungsordnungen basierende Analyse der fachwissenschaftlichen und fachdidaktischen Lerngelegenheiten an insgesamt 28 deutschen Universitäten, die den Studiengang Wirtschaftspädagogik anbieten (Bouley, 2013).

Im vorliegenden Beitrag werden die bisher vorliegenden Erkenntnisse zur Ausprägung von Lerngelegenheiten im Studium der Wirtschaftspädagogik aufgegriffen und in Bezug zu den Möglichkeiten des Erwerbs von Fachwissen und fachdidaktischem Wissen gestellt. Hierzu sind in Abschnitt 2 zunächst die Bedeutung des Fachwissens und fachdidaktischen Wissens für die Qualität des Lehrerhandelns sowie der Zusammenhang zwischen Lerngelegenheiten und Wissen zu diskutieren. In Abschnitt 3 und 4 geht es dann um die Forschungsziele bzw. die Darstellung der Forschungsmethodik. Die empirischen Befunde werden in Abschnitt 5 berichtet. Der Beitrag schließt mit einer Diskussion und einem Ausblick (Abschnitt 6).

1 Das vom Bundesministerium für Bildung und Forschung geförderte Verbundprojekt (Förderkennzeichen: 01PK11003a-c) wird von der Forschergruppe um Seifried, Wuttke und Schmitz an den Universitäten Mannheim, Frankfurt und Darmstadt bearbeitet.

2. Theoretische Grundlagen und Stand der Forschung

2.1 Die Bedeutung fachwissenschaftlichen und fachdidaktischen Wissens für professionelles Lehrerhandeln

Professionelles Wissen gilt als erklärungsmächtiger Faktor für erfolgreiches Lehrerhandeln (z. B. Ball, Thames & Phelps, 2008; Besser & Krauss, 2009; Hill, Rowan & Ball, 2005; Krauss et al., 2008; Sadler, Sonnert, Coyle, Cook-Smith & Miller, 2013). Dabei folgt die Ausdifferenzierung i. d. R. dem Vorschlag von Shulman (1986, 1987), der u. a. zwischen Fachwissen, fachdidaktischem Wissen und allgemein pädagogischem Wissen unterscheidet. Im Mittelpunkt der Diskussion steht seit einiger Zeit insbesondere das fachbezogene Wissen von Lehrkräften, also Fachwissen (Wissen über unterrichtsbezogene Fachinhalte) und fachdidaktisches Wissen (Wissen darüber, wie Fachinhalte verständlich an die Lernenden vermittelt werden können) (z. B. Kleickmann et al., 2014; Neuweg, 2014).

In diesem Zusammenhang ist von besonderem Interesse, ob und wie diese beiden Wissenskomponenten zusammenwirken und inwiefern sie das Handeln der Lehrkräfte und damit in letzter Konsequenz auch den Lernerfolg von Schülerinnen und Schülern beeinflussen. Im Zuge der COACTIV-Studie (Krauss et al., 2008; Kunter et al., 2011) zeigte sich, dass gerade dem fachdidaktischen Wissen von Lehrpersonen Bedeutung zukommt und dass diese Wissenskomponente größeren Einfluss auf Unterrichtsqualität und Lernerfolg ausübt als das Fachwissen. Fachwissen alleine ermöglicht nur in Ausnahmefällen professionelles Lehrerhandeln, da hier der Bezug zu Unterricht und Lernprozessen weitgehend fehlt (Baumert et al., 2010; Kind, 2009; Sullivan, Clarke & Clarke, 2013). Es wird daher als notwendige, aber nicht hinreichende Bedingung für verständnisorientierten Unterricht betrachtet (Baumert et al., 2010). Notwendig deshalb, weil erst auf der Basis fundierten Fachwissens zentrale fachdidaktische Facetten (z. B. die Erarbeitung von Inhalten oder die adäquate Rückmeldung auf Schülerbeiträge) ihre Wirkung entfalten können (Neuweg, 2014).

Der skizzierte Einfluss des Fachwissens auf das fachdidaktische Wissen und Handeln von Lehrkräften ist bislang nicht intensiv beleuchtet worden (Rollnick, Bennett, Rhemtula, Dharsey & Ndlovu, 2008). Üblicherweise wird über Zusammenhänge zwischen den beiden Wissenskomponenten berichtet, die auf empirisch trennbare, aber eng zusammenhängende Konstrukte verweisen. Für Mathematik oder Physik beispielsweise werden hohe latente Korrelationen von bis zu $r = .8$ ermittelt (Krauss et al., 2008; Blömeke, Felbrich & Müller, 2008; Kleickmann et al., 2014; Riese & Reinhold, 2012). Für eher gering strukturierte Domänen wie Englisch oder Deutsch fallen die Werte geringer aus (latente Werte von $r = .4$ bis $r = .6$, siehe Blömeke et al., 2011). Ähnliche (manifeste) Werte werden für die Wirtschaftswissenschaften berichtet ($r = .4$; Kuhn et al., 2014).

2.2 Die Bedeutung von universitären sowie außeruniversitären Lerngelegenheiten für den Wissenserwerb

Die Erklärung von Wissen und Können geschieht i. d. R. mit Blick auf das komplexe Zusammenwirken von (1) Lerngelegenheiten (Lernangeboten) und (2) deren Nutzung durch die Lernenden (Angebot-Nutzungs-Modell, z. B. Ditton, 2000; Helmke, 2009). Im Folgenden geht es also nicht nur um das Bereitstellen von universitären und außeruniversitären Lerngelegenheiten, sondern auch um deren Verwertung durch die Lernenden (bzw. Studierenden).

ad (1) Lerngelegenheiten: Untersuchungen zur Entwicklung der Lehrerprofessionalität belegen die Bedeutung universitärer Lerngelegenheiten (Blömeke, Suhl et al., 2010; Cochran-Smith & Zeichner, 2005; Darge, Schreiber, König & Seifert, 2012; Kunina-Habenicht et al., 2013; Riese & Reinhold, 2012). Für die Wirtschaftswissenschaften deuten die Befunde einer der wenigen Studien (Kuhn et al., 2014) darauf hin, dass einschlägige wirtschaftswissenschaftliche und wirtschaftsdidaktische Lerngelegenheiten das fachdidaktische Wissen beeinflussen. Vor dem Hintergrund der uneinheitlichen Befundlage anderer Domänen (z. B. Mathematik, siehe z. B. Blömeke, Suhl et al., 2010; Kleickmann & Anders, 2011) sind auf dieser Basis kaum Rückschlüsse auf den Zusammenhang von Lerngelegenheiten und Wissen möglich.

ad (2) Nutzung der Lerngelegenheiten: Mit Blick auf die Nutzung von Lerngelegenheiten wird in der Lehr-Lern- und Unterrichtsforschung auf die Bedeutung individueller Dispositionen und insbesondere des Vorwissens verwiesen (Hattie, 2009; König, Tachtsoglou & Seifert, 2012). Im vorliegenden Fall kann außeruniversitär gewonnenes Vorwissen hauptsächlich aus einer vor dem Studium absolvierten kaufmännischen Vollzeitschule (z. B. Wirtschaftsgymnasium, Fachoberschule) oder einer kaufmännischen Berufsausbildung resultieren. Vorwissen kann zudem aus einschlägigen betrieblichen sowie schulischen Praktika im Rechnungswesen resultieren.

Domänenübergreifend lässt sich ein Zusammenhang zwischen Studiendauer und professionellem Wissen ausmachen (Arzi & White, 2008; Blömeke, Suhl et al., 2010; Kennedy, Ahn & Choi, 2008; Schmelzing et al., 2013). Die oben angesprochene Dokumentenanalyse (Bouley, 2013) zeigt jedoch, dass die im vorliegenden Fall relevanten Fachwissensinhalte ausschließlich im ersten Studienjahr des Bachelorstudiums angeboten werden und sich mit Blick auf fachdidaktisch akzentuierte Lehrveranstaltungen kein über alle Universitäten hinweg einheitliches Bild ergibt. Sie verteilen sich unsystematisch über das Bachelor- und Masterstudium. Zudem belegen die von uns durchgeführten Studien zum Fehlerwissen im Rechnungswesen von (angehenden) Lehrkräften an kaufmännischen Schulen (siehe Abschnitt 1), dass sich Bachelor- und Masterstudierende hinsichtlich ihres Kompetenzniveaus nicht signifikant unterscheiden (Türling, 2014; Wuttke & Seifried, 2013). Es ist für die Domäne des Rechnungswesens folglich nicht zu erwarten, dass fortgeschrittene Studierende bezüglich des Fachwissens und des fachdidaktischen Wissens im Vergleich zu Studienanfängern Vorteile aufweisen.

3. Forschungsfragen

(1) Wie hängen Fachwissen und fachdidaktisches Wissen im Bereich des Rechnungswesens zusammen?
Vor dem Hintergrund der oben skizzierten Befunde zum Zusammenhang zwischen Fachwissen und fachdidaktischem Wissen erwarten wir auch für das Rechnungswesen einen positiven Zusammenhang (Hypothese 1).

(2) Inwiefern beeinflussen universitäre und außeruniversitäre Lerngelegenheiten Fachwissen und fachdidaktisches Wissen?
Die Annahme der prädiktiven Kraft von universitären und außeruniversitären Lerngelegenheiten hinsichtlich der Ausprägung des professionellen Wissens ist empirisch untermauert. Daraus resultiert für die vorliegende Untersuchung die Erwartung, dass sich der Besuch von universitären fachwissenschaftlichen Veranstaltungen positiv auf das Fachwissen auswirken sollte (Hypothese 2). An dieser Stelle ist anzumerken, dass für Studierende der Wirtschaftspädagogik eine grundlegende Veranstaltung zum Rechnungswesen obligatorisch ist. Interessant ist daher insbesondere, ob die Nutzung von darüber hinaus im Wahl(pflicht)bereich angesiedelten einschlägigen Veranstaltungen einen Zugewinn verspricht. Ferner wird angenommen, dass sich außeruniversitäre Lerngelegenheiten wie Betriebspraktika, der Abschluss einer kaufmännischen Berufsausbildung und der Besuch einer kaufmännischen Vollzeitschule positiv auf das Fachwissen auswirken (Hypothese 3). Analog dazu wird erwartet, dass der Besuch von fachwissenschaftlichen und fachdidaktischen universitären Veranstaltungen (Hypothese 4) fachdidaktisches Wissen positiv beeinflusst. Auch ein Einfluss von außeruniversitären Lerngelegenheiten wie Betriebs- und Schulpraktika sowie der Abschluss einer kaufmännischen Berufsausbildung und der Besuch einer kaufmännischen Vollzeitschule erscheinen plausibel (Hypothese 5).

4. Methode und Stichprobe

4.1 Stichprobe

In die Analyse gehen die Angaben von insgesamt N = 1 152 Studierenden (männlich: 395, weiblich: 751, keine Angabe: 6) an 24 deutschen Hochschulstandorten, die den Studiengang Wirtschaftspädagogik anbieten, ein. Die Mehrheit der Studierenden (590 Personen) befand sich im Bachelor-, weitere 552 im Masterprogramm (keine Angabe: 10). Der Teilstichprobenumfang pro Hochschule variiert zwischen 4 und 148. Da kein Vergleich der Hochschulen angestrebt wird, erscheint dieser Umstand vertretbar. Die Studierenden waren im Durchschnitt 24.8 Jahre alt (SD = 3.51). 761 Probanden (keine Angabe: 5) berichten über einschlägige Vorerfahrung im Berufsbildungssystem: 489 Studierende besuchten eine kaufmännische Vollzeitschule und knapp die Hälfte der Probanden absolvierte eine kaufmännische Berufsausbildung. 159 Studierende er-

Erfahrungen im Berufsbildungssystem	n *(keine Angabe)*
Ohne kaufmännisch-berufsbildende Vorbildung	386 *(5)*
Kaufmännische Vollzeitschule	489 *(6)*
Davon Abitur am Wirtschaftsgymnasium	463 *(11)*
Kaufmännische Berufsausbildung	535 *(28)*
Kaufmännische Vollzeitschule und kaufmännische Berufsausbildung	263 *(28)*

Tab. 1: Berufsbildungsbiografie der Probanden (Mehrfachantworten möglich)

warben im Rahmen eines Praktikums bereits praktische Erfahrungen im Rechnungswesen. Schulische Praktika schließlich können 614 Probanden nachweisen (keine Angaben: 61).

4.2 Testinstrument

Fachwissen und fachdidaktisches Wissen wurde mit dem „Test zum fachlichen und fachdidaktischen Wissen im Rechnungswesen“ erhoben (zur Testentwicklung siehe Berger et al., 2013; Berger et al., im Druck; Mindnich, Berger & Fritsch, 2013). Dieser besteht aus insgesamt 49 Wissensitems, wovon 35 in geschlossenem und 14 in offenem Antwortformat präsentiert wurden. Das Testinstrument berücksichtigt drei zentrale Facetten fachdidaktischen Wissens: (1) Wissen über Schülerkognitionen und typische Schülerfehler, (2) Wissen über das Potenzial von Aufgaben, (3) Wissen über das Zugänglichmachen und Erklären von Inhalten (siehe Shulman, 1986; Krauss et al., 2008) sowie die bedeutsamsten Lerninhaltsbereiche des externen Rechnungswesens (siehe Berger et al., 2013): (1) Aufgaben und (Rechts-)Grundlagen, (2) System der Doppik, (3) Beschaffungs- und Absatzprozesse. Die Inhaltsbereiche wurden auf Basis von Lehrplan- und Curriculumanalysen sowie Experteninterviews bestimmt (siehe Berger et al., 2013), um die Inhaltsvalidität zu gewährleisten (zur Notwendigkeit des Konzepts der Inhaltsvalidität siehe Jenßen, Dunekacke & Blömeke, 2015, in diesem Beiheft). Die Items verteilen sich gleichmäßig über die Lerninhaltsbereiche sowie die fachdidaktischen Facetten.

Dem Wissenstest wurde in Analogie zur Vorgehensweise in PISA, TIMSS oder COACTIV ein Multi-Matrix-Design in Form von Youden Squares (Frey, Hartig & Rupp, 2009) zugrunde gelegt (sieben Testhefte). Angesichts der Vielzahl an Items und einer gleichzeitigen Restriktion der Testzeit auf 40 Minuten konnte so gewährleistet werden, dass nicht jeder Proband alle 49, sondern lediglich eine Auswahl von 28 Items zu bearbeiten hatte. Die Items wurden in einem ersten Zugriff dichotom kodiert, eine Auswertung unter Nutzung von Partial Credits ist jedoch grundsätzlich möglich und für weiterführende Analysen auch vorgesehen. Items, die dem Probanden aufgrund des Testheftdesigns nicht vorlagen, wurden als fehlend berücksichtigt (missing by design, Rost, 2004).

			Unterschied		
	Deviance	**Parameter**	**Deviance**	**Parameter**	**p**
1-dimensional	40.490,81	50	57.71	2	< .001
2-dimensional	40.433,10	52			

Tab. 2: Gegenüberstellung von ein- und zweidimensionalem Wissensmodell

Die Überprüfung der Dimensionalität des Testinstrumentes (Schnick-Vollmer et al., im Druck) ergab, dass ein zweidimensionales Modell einen besseren Modellfit zeigt als ein entsprechendes eindimensionales Modell ($\chi^2 = 28.86$, df = 2, $p < .001$) (Tab. 2). Fachwissen und fachdidaktisches Wissen können somit als zwei getrennte Skalen angesehen werden.

Exploratorische Faktorenanalysen zeigen dann, dass sowohl das Fachwissen als auch das fachdidaktische Wissen als eindimensionale Konstrukte aufgefasst werden können, sodass sich Analysen auf die zwei empirisch trennbaren Wissensdimensionen beschränken lassen. Überdies sprechen die Daten für ein Rasch-Modell. Die Reliabilitäten für die Fachwissens- und die Fachdidaktikskala liegen jeweils im akzeptablen Bereich von .63 bzw. .62 (Bühner, 2011). Ferner nehmen die gewichteten MNSQ-Kennzahlen zur Bestimmung der Itemqualität gute Werte an ($0.93 \leq \text{MNSQ} \leq 1.07$). Die Auswertung der Itemschwierigkeit zeigt, dass die Items der Dimension Fachdidaktik mit der Personenfähigkeit gut übereinstimmen, für Items der Dimension des Fachwissens kann eine überwiegende Übereinstimmung festgestellt werden (Schnick-Vollmer et al., im Druck).

Die wahrgenommenen Lerngelegenheiten der Probanden wurden mittels eines biografischen Fragebogens erfasst. Dokumentenanalysen zeigen, dass an sämtlichen Hochschulstandorten eine vergleichbare Anzahl an verpflichtenden Rechnungswesenveranstaltungen zu belegen ist, die Summe der Lerngelegenheiten aus dem Wahlbereich dagegen stark variiert (Bouley, 2013). Daher wurde zur Bestimmung universitärer fachwissenschaftlicher Lerngelegenheiten die Anzahl an Veranstaltungen des externen Rechnungswesens abgefragt, die die Studierenden über den Pflichtkanon hinaus besuchten. Zur Bestimmung außeruniversitärer Lerngelegenheiten wurden neben betrieblichen und schulischen Praktika auch berufsschulische Erfahrungen (Abschluss einer beruflichen Ausbildung und Besuch einer kaufmännischen Vollzeitschule) erfasst (jeweils in Form einer Dummyvariablen). Ferner wurde nach dem Besuch einschlägiger fachdidaktischer Veranstaltungen gefragt (Dummyvariablen). Es böte sich möglicherweise auch an, die ECTS-Summe abzufragen, jedoch zeigten Erfahrungen aus dem Pretest, dass die Probanden sich kaum in der Lage sehen, diesbezüglich präzise Angaben zu machen. Da nur wenige Probanden fachdidaktische Veranstaltungen im Rechnungswesen besuchten (n = 94), wurden in der Analyse ebenfalls thematisch breiter angelegte fachdidaktische Veranstaltungen zur allgemeinen Wirtschaftslehre (n = 346) berücksichtigt.

4.3 Auswertungsmethoden

Als Maß für das Fachwissen und das fachdidaktische Wissen der Teilnehmenden werden die auf Basis der Item-Response-Theorie ermittelten Personenparameter (Ermittlung von WLE-Schätzer mittels Conquest) verwendet. Die durchgeführten Korrelations- und Regressionsanalysen wurden mit IBM SPSS Statistics Version 22 berechnet. Die Prüfung der Voraussetzungen für die Durchführung von Regressionsanalysen liefert zufriedenstellende Ergebnisse: Durban-Watson-, Toleranz- sowie VIF-Werte liegen jeweils im akzeptablen Bereich. Ferner konnte kein Hinweis auf Homoskedastizität sowie Linearität gefunden werden.

5. Empirische Befunde

5.1 Zusammenhang zwischen Fachwissen und fachdidaktischem Wissen

Die Berechnung des Zusammenhangs erfolgt auf Basis einer Pearson-Korrelation. Dabei zeigt sich eine signifikante latente Korrelation zwischen Fachwissen und fachdidaktischem Wissen mittlerer Stärke ($r = .45$, $p \leq .001$). Hypothese 1 kann somit bestätigt werden.

5.2 Einfluss universitärer und außeruniversitärer Lerngelegenheiten auf das Fachwissen

Hypothese 2 und 3 postulieren einen Einfluss universitärer (über die Grundlagenveranstaltung hinaus) und außeruniversitärer fachwissenschaftlicher Lerngelegenheiten auf das Fachwissen. Da in Abschnitt 5.1 gezeigt werden konnte, dass Fachwissen und fachdidaktisches Wissen korrelieren, wurden die nachfolgenden Regressionsanalysen zum Einfluss universitärer und außeruniversitärer Lerngelegenheiten auf das Fachwissen jeweils unter Kontrolle des fachdidaktischen Wissens durchgeführt und somit um dessen Einfluss bereinigt.

In einem ersten Schritt wurde der Einfluss fachwissenschaftlicher universitärer (Zusatz-)Veranstaltungen im Rechnungswesen auf das Fachwissen untersucht (Modell 1). Wie Tabelle 3 zeigt, tragen diese nicht signifikant zum Fachwissen angehender Lehrpersonen bei. Modell 2 beinhaltet den Einfluss außeruniversitärer Lerngelegenheiten. Hier zeigt sich, dass sowohl Betriebspraktika als auch der Abschluss einer kaufmännischen Ausbildung und der Besuch einer kaufmännischen Vollzeitschule signifikant positiv auf Fachwissen wirken. Modell 3 berücksichtigt sowohl universitäre als auch außeruniversitäre Lerngelegenheiten. Hier kann der Einfluss außeruniversitärer Lerngelegenheiten bestätigt werden. Folglich ist Hypothese 2 abzulehnen; Hypothese 3 dagegen wird unterstützt.

	Fachwissen					
	Modell 1		**Modell 2**		**Modell 3**	
	B	**β**	**B**	**β**	**B**	**β**
Fachdidaktisches Wissen	.746	.454***	.633	.385***	.632	.385***
Universitäre Lerngelegenheiten						
Zusätzlich besuchte fachwissenschaftliche Veranstaltungen (n = 358)	.046	.053	–	–	.030	.034
Außeruniversitäre Lerngelegenheiten						
Betriebliches Praktikum im Rechnungswesen (n = 159)			.187	.057*	.178	.054*
Kfm. Ausbildung (n = 535)			.405	.181***	.402	.180***
Kaufmännische Vollzeitschule (n = 489)			.216	.095***	.210	.093***
R^2	.211		.254		.255	

* p ≤ .05; ** p ≤ .01; *** p ≤ .001; B = Regressionskoeffizient, β = standardisierter Regressionskoeffizient

Tab. 3: Einfluss von universitären und außeruniversitären Lerngelegenheiten auf das Fachwissen unter Kontrolle des fachdidaktischen Wissens (Hypothese 2 und 3)

5.3 Einfluss universitärer und außeruniversitärer Lerngelegenheiten auf das fachdidaktische Wissen

In Hypothese 4 und 5 geht es um den Einfluss universitärer und außeruniversitärer Lerngelegenheiten auf das fachdidaktische Wissen. Auch hier wird die jeweils andere Wissenskomponente (in diesem Fall das Fachwissen) zur Kontrolle einbezogen. Die Ergebnisse sind in Tabelle 4 dargestellt.

Fachdidaktische Veranstaltungen üben einen signifikanten, wenn auch geringen Einfluss auf das fachdidaktische Wissen aus (Modell 1). Für außerschulische Lerngelegenheiten kann ein Einfluss des Abschlusses einer kaufmännischen Ausbildung nachgewiesen werden (Modell 2). In Modell 3 werden sowohl universitäre als auch außeruniversitäre Lerngelegenheiten berücksichtigt. Hier zeigt sich, dass nur noch der Besuch einer kaufmännischen Ausbildung Auswirkungen auf fachdidaktisches Wissen hat. Für die weiteren Variablen wird kein Einfluss nachgewiesen. Hypothese 4 ist demnach abzulehnen. Hypothese 5 kann lediglich in Bezug auf die kaufmännische Ausbildung unterstützt werden.

	Fachdidaktisches Wissen					
	Modell 1		Modell 2		Modell 3	
	B	β	B	β	B	β
Fachwissen	.272	.451***	.235	.390***	.233	.387***
Universitäre Lerngelegenheiten						
Fachdidaktische Veranstaltungen (n = 406)	.097	.070*	–	–	.079	.057
Zusätzlich besuchte fachwissenschaftliche Veranstaltungen (n = 358)	.005	.009	–	–	.001	.002
Außeruniversitäre Lerngelegenheiten						
Schulpraktikum (n = 614)			.021	.016	−.006	−.004
Betriebliches Praktikum im Rechnungswesen (n = 159)			.068	.034	.065	.033
Kfm. Ausbildung (n = 535)			.251	.185***	.250	.184***
Kaufmännische Vollzeitschule (n = 489)			.040	.029	.036	.026
R^2	.215		.245		.248	

* p ≤ .05; ** p ≤ .01; *** p ≤ .001; B = Regressionskoeffizient, β = standardisierter Regressionskoeffizient

Tab. 4: Einfluss universitärer und außeruniversitärer Lerngelegenheiten auf das fachdidaktische Wissen unter Kontrolle des Fachwissens (Hypothese 4 und 5)

6. Diskussion und Fazit

Der identifizierte mittlere latente Zusammenhang zwischen fachwissenschaftlichem und fachdidaktischem Wissen harmoniert mit vorliegenden Befunden aus dem Bereich der Wirtschaftswissenschaften (Kuhn et al., 2014). Im Vergleich zu anderen Domänen fällt der Zusammenhang zwischen den Wissenskomponenten jedoch geringer aus. Dies könnte auf die Operationalisierung der beiden Konstrukte zurückgeführt werden (siehe dazu auch Neuweg, 2014): Der eingesetzte Fachwissenstest beschränkt sich auf grundlegende Kenntnisse des Rechnungswesens, während sich der Fachdidaktiktest auf ein komplexeres anforderungsorientiertes Verständnis bezieht. Das Ergebnis mittlerer Zusammenhänge entspricht damit weitgehend den Erwartungen.

Betrachtet man den Einfluss von Lerngelegenheiten, so zeigt sich, dass unter Kontrolle der jeweils anderen Wissensdimension universitäre Schwerpunktsetzungen im Rechnungswesen erstaunlicherweise keinen Erklärungsbeitrag zum fachwissenschaftlichen und fachdidaktischen Wissensstand leisten. Außerschulische Lerngelegenheiten weisen dagegen einen Einfluss auf die Ausprägung von Fachwissen auf. Das gilt auch – zumindest mit Blick auf die kaufmännische Ausbildung – für das fachdidaktische Wissen. Auch wenn die Befunde zu den außerschulischen Lerngelegenheiten signifikant

sind, klären die Lerngelegenheiten tatsächlich nur wenig Varianz im Fachwissen und im fachdidaktischen Wissen auf.

Dieses Ergebnis, das zuerst einmal nahezulegen scheint, dass weder universitäre Zusatzveranstaltungen im Rechnungswesen noch außeruniversitär besuchte einschlägige kaufmännisch-berufsbildende Veranstaltungen eine nennenswerte Rolle für die Entwicklung von Fachwissen und fachdidaktischem Wissen spielen, wirft einige zentrale Fragen auf:

a) Grundsätzlich stellt sich die Frage, bei welcher Gelegenheit Fachwissen und fachdidaktisches Wissen erworben werden, wenn nicht in einschlägigen außeruniversitären oder universitären Kontexten. Allerdings liegen möglicherweise die Probleme eher bei den eingeschränkten Möglichkeiten der Erfassung der Lerngelegenheiten bzw. der Testleistung.
b) Bei den Lerngelegenheiten ist zu fragen, wie aussagekräftig sie erhoben wurden. Beim bisher gewählten Zugriff wurde lediglich deren Anzahl relativ grob erfasst. Aussagen über Qualität, tatsächlichen Inhalt oder zeitlichen Umfang sind nur eingeschränkt möglich. Auch die Hinzunahme von Kreditpunkten oder Semesterwochenstunden ist nur bedingt zielführend, da diese Größen ebenfalls wenige Rückschlüsse auf die Qualität der Lerngelegenheiten zulassen. Hier sind weiterführende Analysen vonnöten, um fundierte Aussagen zur Wirksamkeit universitärer Lehrerbildung treffen zu können.
c) Im Hinblick auf das Testinstrument ist insbesondere die Frage der curricularen Validität in den Blick zu nehmen. Bei der Entwicklung der Testitems wurde outputorientiert vorgegangen. Lehrplan- und Schulbuchanalysen sowie Experteninterviews sollten sicherstellen, dass die Testitems Anforderungen widerspiegeln, die an angehende Lehrkräfte gestellt werden und die sie deshalb nach Abschluss des Studiums beherrschen müssten. Wie gut die – eher heterogene – Ausbildung an den einzelnen Standorten der Wirtschaftspädagogik auf diese Anforderungen überhaupt (einheitlich) vorbereitet, ist allerdings offen. Gelingt dies nicht, ist der fehlende Beitrag universitärer Lerngelegenheiten zum Aufbau von mit diesem Test erfasstem Fachwissen und fachdidaktischem Wissen kaum erstaunlich.

Bezogen auf die Datenauswertung sind weitere Analyseschritte zu gehen. So wären beispielsweise Mehrebenenanalysen zur Kontrolle des Einflusses des universitären Standortes von Interesse. Auch Strukturgleichungsmodelle sollten in einem weiteren Schritt berücksichtigt werden, um eine messfehlerbereinigte Schätzung der Korrelationen zwischen Fachwissen und fachdidaktischem Wissen gewährleisten zu können. Ferner erlauben Strukturgleichungsmodelle eine weitaus differenziertere Berücksichtigung aller Einflussgrößen als klassische Regressionsmodelle. Darüber hinaus wird in weiterführenden Schritten statt der bislang durchgeführten dichotomen Kodierung der Testitems eine polytome Kodierung (Partial Credits, s. o.) durchgeführt. In einem weiteren Schritt werden dann auch Plausible Values auf der Basis von diversen Hintergrundmodellen berechnet.

Literatur

Abell, S. K. (2007). Research on science teacher knowledge. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of research on science education* (S. 1105–1149). Mahwah: Erlbaum.

Achtenhagen, F. (1996). Entwicklung ökonomischer Kompetenz als Zielkategorie des Rechnungswesenunterrichts. In P. Preiß & T. Tramm (Hrsg.), *Rechnungswesenunterricht und ökonomisches Denken* (S. 22–44). Wiesbaden: Gabler.

Arzi, H., & White, R. T. (2008). Changes in teachers' knowledge of subject matter: A 17-year longitudinal study. *Science Education, 9*, 221–251.

Ball, D. L., Lubienski, S., & Mewborn, D. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Hrsg.), *Handbook of research on teaching* (4. Aufl., S. 433–456). Washington, D.C.: American Educational Research Association.

Ball, D. L., Thames, M. H., & Phelps, G. C. (2008). Content Knowledge for Teaching: What Makes It Special? *Journal of Teacher Education, 59*(5), 389–407.

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

Berger, S., Bouley, F., Fritsch, S., Krille, C., Seifried, J., & Wuttke, E. (im Druck). Fachwissen und fachdidaktisches Wissen im wirtschaftspädagogischen Studium – Entwicklung eines Testinstruments und erste empirische Befunde. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Hrsg.), *Kompetenzen von Lehramtsstudierenden und angehenden ErzieherInnen.* Bad Heilbrunn: Klinkhardt.

Berger, S., Fritsch, S., Seifried, J., Bouley, F., Mindnich, A., Wuttke, E., Schnick-Vollmer, K., & Schmitz, B. (2013). Entwicklung eines Testinstruments zur Erfassung des fachlichen und fachdidaktischen Wissens von Studierenden der Wirtschaftspädagogik – Erste Erfahrungen und Befunde. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften* (Lehrerbildung auf dem Prüfstand, Sonderheft, S. 93–107). Landau: Verlag Empirische Pädagogik.

Besser, M., & Krauss, S. (2009). Zur Professionalität als Expertise. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 71–82). Weinheim/Basel: Beltz.

Blömeke, S., Bremerich-Vos, A., Haudeck, H., Kaiser, G., Nold, G., Schwippert, K., & Willenberg, H. (Hrsg.) (2011). *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen – Erste Ergebnisse aus TEDS-LT*. Münster: Waxmann.

Blömeke, S., Felbrich, A., & Müller, C. (2008). Erziehungswissenschaftliches Wissen am Ende der Lehrerausbildung. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematik-Studierender und -Referendare – Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung* (S. 195–217). Münster: Waxmann.

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

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

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., & Fege, J. (Hrsg.) (2013). *Modeling and Measuring Competencies in Higher Education: Task and Challenges*. Rotterdam: Sense Publishers.

BMBF Bundesministerium für Bildung und Forschung (2012). *Berufliche Kompetenzen sichtbar machen. Die Forschungsinitiative ASCOT*. Bonn: BMBF.

Bouley, F. (2013). *Die Lerngelegenheiten in der Domäne Rechnungswesen und seiner Fachdidaktik im wirtschaftspädagogischen Studium in deutschen Hochschulen*. Unveröffentlichte Masterarbeit am Lehrstuhl für Wirtschaftspädagogik, insb. Lehr-Lernforschung der Goethe-Universität Frankfurt.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.

Cochran-Smith, M., & Zeichner, K. M. (Hrsg.) (2005). *Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education*. Washington, D.C.: American Educational Research Association.

Darge, K., Schreiber, M., König, J., & Seifert, A. (2012). Lerngelegenheiten im erziehungswissenschaftlichen Studium. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* (S. 87–118). Münster: Waxmann.

Ditton, H. (2000). Qualitätskontrolle und -sicherung in Schule und Unterricht – ein Überblick zum Stand der empirischen Forschung. *Zeitschrift für Pädagogik, 41. Beiheft*, 73–92.

Döbrich, P., Klemm, K., Knauss, G., & Lange, H. (2003). *Ausbildung, Einstellung und Förderung von Lehrerinnen und Lehrern (OECD-Lehrerstudie). Ergänzende Hinweise zu dem Nationalen Hintergrundbericht (CBR) für die Bundesrepublik Deutschland. (Bericht für die Kultusministerkonferenz)*. http://www.oecd.org/dataoecd/55/61/3107 6280.pdf [23. 04. 2014].

Eurydice (2003). *Der Lehrerberuf in Europa: Profil, Tendenzen und Anliegen. Bericht III: Beschäftigungsbedingungen und Gehälter. Allgemein bildender Sekundarbereich I*. Brüssel: Eurydice.

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

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53.

Halim, L., & Meerah, S. M. (2002). Science Trainee Teachers' Pedagogical Content Knowledge and its Influence on Physics Teaching. *Research in Science & Technological Education, 20*(2), 215–225.

Hascher, T. (2011). Forschung zur Wirksamkeit der Lehrerbildung. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 418–440). Münster: Waxmann.

Hattie, J. A. C. (2009). *Visible Learning: A synthesis of 800+ meta-analyses on achievement*. London: Routledge.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (2. Aufl.). Seelze: Klett-Kallmeyer.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for teaching on Student Achievement. *American Educational Research Journal, 42*, 371–406.

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik, 61. Beiheft*, 11–31.

Kennedy, M. M., Ahn, S., & Choi, J. (2008). The value added by teacher education. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Hrsg.), *Handbook of research on teacher education* (3. Aufl., S. 1249–1273). London: Routledge.

Kind, V. (2009). Pedagogical content knowledge in science education: Perspectives and potential for progress. *Studies in Science Education, 45*, 169–204.

Kleickmann, T., & Anders, Y. (2011). Lernen an der Universität. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften – Ergebnisse des Forschungsprogramms COACTIV* (S. 305–315). Münster: Waxmann.

Kleickmann, T., Großschedl, J., Harms, U., Heinze, A., Herzog, S., Hohenstein, F., Köller, O., Kröger, J., Lindmeier, A., Loch, C., Mahler, D., Möller, J., Neumann, K., Parchmann, I., Steffensky, M., Taskin, V., & Zimmermann, F. (2014). Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer – Testentwicklung im Rahmen des Projekts KiL. *Unterrichtswissenschaft, 42*(3), 280–288.

König, J., Tachtsoglou, S., & Seifert, A. (2012). Individuelle Voraussetzungen, Lerngelegenheiten und der Erwerb von pädagogischem Professionswissen. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* (S. 243–283). Münster: Waxmann.

Krauss, S., Neubrand, M., Blum, W., Baumert, J., Brunner, M., Kunter, M., & Jordan, A. (2008). Die Untersuchung des professionellen Wissens deutscher Mathematik-Lehrerinnen und -Lehrer im Rahmen der COACTIV-Studie. *Journal für Mathematikdidaktik, 29*(3/4), 223–258.

Kuhn, C., Happ, R., Zlatkin-Troitschanskaia, O., Beck, K., Förster, M., & Preuße, D. (2014). Kompetenzentwicklung angehender Lehrkräfte im kaufmännisch-verwaltenden Bereich – Erfassung und Zusammenhänge von Fachwissen und fachdidaktischem Wissen. *Zeitschrift für Erziehungswissenschaft, 17*(1), 149–167.

Kunina-Habenicht, O., Schulze-Stocker, F., Kunter, M., Baumert, J., Leutner, D., Förster, D., Lohse-Bossenz, H., & Terhart, E. (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., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften – Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Mindnich, A., Berger, S., & Fritsch, S. (2013). Modellierung des fachlichen und fachdidaktischen Wissens von Lehrkräften im Rechnungswesen – Überlegungen zur Konstruktion eines Testinstruments. In U. Faßhauer, B. Fürstenau & E. Wuttke (Hrsg.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung 2013* (S. 61–72). Opladen: Budrich.

Neuweg, G. H. (2014). Das Wissen der Wissensvermittler. Problemstellungen, Befunde und Perspektiven der Forschung zum Lehrerwissen. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (2. überarb. u. erw. Aufl., S. 583–614). Münster/New York: Waxmann.

OECD (2005). *Teachers matter. Attracting, developing and retaining effective teachers*. Paris: OECD.

Oser, F., & Oelkers, J. (Hrsg.) (2001). *Die Wirksamkeit der Lehrerbildungssysteme. Von der Allrounderbildung zur Ausbildung professioneller Standards*. Zürich: Rüegger.

Preiß, P. (2000). Der Rechnungswesenunterricht als Beitrag zum Verständnis ökonomischer Zusammenhänge und wirtschaftlicher Entscheidungen. In Bundesverband der Lehrer an Wirtschaftsschulen (Hrsg.), *Funktionswandel des Rechnungswesens: Von der Dokumentation zur Steuerung* (Heft 44 der Sonderschriftreihen des VLW, S. 7–29).

Riese, J., & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen. *Zeitschrift für Erziehungswissenschaft, 15*(1), 111–143.

Rollnick, M., Bennett, J., Rhemtula, M., Dharsey, N., & Ndlovu, T. (2008). The place of subject matter knowledge in pedagogical content knowledge: A case study of South African teachers teaching the amount of substance and chemical equilibrium. *International Journal of Science Education, 30*(10), 1365–1387.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2., vollst. überarb. u. erw. Aufl.). Bern: Huber.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013) Student learning in middle school science classrooms. *American Educational Research Journal, 50*, 1020–1049.

Schaefers, C. (2002). Forschung zur Lehrerausbildung in Deutschland – eine bilanzierende Übersicht der neueren empirischen Studien. *Schweizerische Zeitschrift für Bildungswissenschaften, 24*(1), 65–90.

Schmelzing, S., van Driel, J. H., Jüttner, M., Brandenbusch, S., Sandman, A., & Neuhaus, B. J. (2013). Development, Evaluation, and Validation of a paper-and-pencil test for measuring two components of biology teachers' pedagogical content knowledge concerning the „Cardiovascular system". *International Journal of Science and Mathematics Education, 11*(6), 1369–1390.

Schnick-Vollmer, K., Berger, S., Bouley, F., Fritsch, S., Schmitz, B., Seifried, J., & Wuttke, E. (im Druck). Modeling Competencies of Prospective Teachers in Business and Economics Education: Professional Knowledge in Accounting. *Zeitschrift für Psychologie*.

Seeber, S., Nickolaus, R., Winther, E., Achtenhagen, F., Breuer, K., Frank, I., Lehmann, R., Spöttl, G., Straka, G., Walden, G., Weiß, R., & Zöller, A. (2010). *Kompetenzdiagnostik in der Berufsbildung. Begründung und Ausgestaltung eines Forschungsprogramms*. In Beilage BWP. Berufsbildung in Wissenschaft und Praxis 1/2010. Bonn.

Seifried, J., Türling, J. M., & Wuttke, E. (2010). Professionelles Lehrerhandeln – Schülerfehler erkennen und für Lernprozesse nutzen. In J. Warwas & D. Sembill (Hrsg.), *Schulleitung zwischen Effizienzkriterien und Sinnfragen* (S. 137–156). Baltmannsweiler: Schneider Verlag Hohengehren.

Shulman, L. S. (1986). Those who understand. Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Sullivan, T., Clarke, D., & Clarke, B. (2013). *Teaching with Tasks for Effective Mathematics Learning*. New York: Springer.

Thanheiser, E. (2009). Preservice elementary school teachers' conception of multidigit whole numbers. *Journal for Research in Mathematics Education, 40*, 251–281.

Türling, J. M. (2014). *Die professionelle Fehlerkompetenz von (angehenden) Lehrkräften – eine empirische Untersuchung im Rechnungswesenunterricht*. Wiesbaden: Springer VS.

Wuttke, E., & Seifried, J. (2013). Diagnostic Competence of (Prospective) Teachers in Vocational Education: An Analysis of Error Identification in Accounting Lessons. In K. Beck & O. Zlatkin-Troitschanskaia (Hrsg.), *From Diagnostics to Learning Success. Proceedings in Vocational Education and Training* (S. 225–240). Rotterdam: Sense Publishers.

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., Hansen, M., & Happ, R. (2013). Modellierung und Erfassung der wirtschaftswissenschaftlichen Fachkompetenz bei Studierenden im deutschen Hochschulbereich. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften* (Lehrerbildung auf dem Prüfstand, Sonderheft, S. 108–133). Landau: Verlag Empirische Pädagogik.

Abstract: There is evidence that content knowledge and pedagogical content knowledge significantly affect teaching quality. Therefore, teacher education should provide learning opportunities that allow the construction and development of these knowledge facets. In our paper, we discuss the influence of learning opportunities at university and those that are content-relevant but not university related, taking the education of prospective teachers at vocational schools as example. Results show a positive influence of learning opportunities that are not university related (especially commercial education) on content knowledge and pedagogical content knowledge. Learning opportunities at university play a minor role.

Keywords: Content Knowledge, Pedagogical Content Knowledge, Learning Opportunities, Teacher Education, Economic Education

Anschrift der Autor(inn)en

M.Sc. Franziska Bouley, Goethe-Universität Frankfurt am Main, Fachbereich Wirtschaftswissenschaften, Professur für Wirtschaftspädagogik, insb. Lehr-Lern-Forschung, Grüneburgplatz 1, 60323 Frankfurt am Main, Deutschland
E-Mail: bouley@econ.uni-frankfurt.de

M.A. Stefanie Berger, Universität Mannheim, Fakultät für Betriebswirtschaftslehre, Lehrstuhl für Wirtschaftspädagogik II, L4, 1, 68161 Mannheim, Deutschland
E-Mail: stefanie.berger@bwl.uni-mannheim.de

M.A. Sabine Fritsch, Universität Mannheim, Fakultät für Betriebswirtschaftslehre, Lehrstuhl für Wirtschaftspädagogik II, L4, 1, 68161 Mannheim, Deutschland
E-Mail: fritsch@bwl.uni-mannheim.de

Prof. Dr. Eveline Wuttke, Goethe-Universität Frankfurt am Main, Fachbereich Wirtschaftswissenschaften, Professur für Wirtschaftspädagogik, insb. Lehr-Lern-Forschung, Grüneburgplatz 1, 60323 Frankfurt am Main, Deutschland
E-Mail: wuttke@em.uni-frankfurt.de

Prof. Dr. Jürgen Seifried, Universität Mannheim, Fakultät für Betriebswirtschaftslehre, Lehrstuhl für Wirtschaftspädagogik II, L4, 1, 68161 Mannheim, Deutschland
E-Mail: seifried@bwl.uni-mannheim.de

Dipl.-Psych. Kathleen Schnick-Vollmer, Technische Universität Darmstadt, Institut für Psychologie, Arbeitsgruppe Pädagogische Psychologie, Alexanderstraße 10, 64283 Darmstadt, Deutschland
E-Mail: schnick@psychologie.tu-darmstadt.de

Prof. Dr. Bernhard Schmitz, Technische Universität Darmstadt, Institut für Psychologie, Arbeitsgruppe Pädagogische Psychologie, Alexanderstraße 10, 64283 Darmstadt, Deutschland
E-Mail: schmitz@psychologie.tu-darmstadt.de

Olga Zlatkin-Troitschanskaia/Manuel Förster/Susanne Schmidt/
Sebastian Brückner/Klaus Beck

Erwerb wirtschaftswissenschaftlicher Fachkompetenz im Studium

Eine mehrebenenanalytische Betrachtung von hochschulischen und individuellen Einflussfaktoren

Zusammenfassung: Für die wirtschaftswissenschaftlichen Studienfächer liegen bislang nur wenige Erkenntnisse zum Einfluss der hochschulischen Lernangebote auf den Fachkompetenzerwerb im Studium vor. Aufbauend auf den Daten aus dem WiwiKom-Projekt mit bundesweiten Erhebungen an insgesamt 33 Hochschulen mit wirtschaftswissenschaftlichen Fakultäten erfolgt im vorliegenden Beitrag eine Analyse, inwieweit die wirtschaftswissenschaftliche Fachkompetenz von Studierenden durch hochschulische und individuelle Einflussfaktoren erklärt wird. Methodisch wird eine Mehrebenenanalyse verwendet, die die hierarchische Struktur des Hochschulsektors angemessen berücksichtigt und zugleich eine unkonfundierte Schätzung der Einflüsse hochschulischer Lernangebote erlaubt. Die Ergebnisse weisen auf bedeutsame interhochschulische und interindividuelle Unterschiede hin, die abschließend kritisch diskutiert werden.

Schlagworte: Fachkompetenz, Wirtschaftswissenschaften, Mixed Methods, Mehrebenenanalyse, Validität

1. Einleitung und Problemstellung

Seit der Bologna-Vereinbarung müssen Studiengänge und somit auch die Lehrpraxis an den Universitäten kompetenzorientiert gestaltet werden. Damit erfolgt bei der Konzeption der Lehre ein Paradigmenwechsel hin zur Fokussierung auf die in der Hochschullehre zu vermittelnden *fachlichen und fachübergreifenden* Kompetenzen von Studierenden. In den Vordergrund rückt die Frage, welche Kompetenzen Studierende im Laufe einer Lehrveranstaltung (LV) sowie ihres gesamten Studiums erwerben bzw. am Ende der Hochschulausbildung erworben haben sollen. Über diese inhaltlichen Fragen hinaus verlangt die Erfassung der akademisch vermittelten Kompetenzen bei Studierenden und Hochschulabsolventen vorab die Entwicklung von angemessenen Modellen sowie reliablen und validen Messinstrumenten.

Betrachtet man den aktuellen internationalen Forschungsstand zur Kompetenzmodellierung und Kompetenzerfassung in verschiedenen Studiendomänen, so wird man v. a. im hochschulischen Lehrerbildungsbereich für allgemeinbildende Schulfächer wie Mathematik, Naturwissenschaften und Sprachen fündig (Kuhn & Zlatkin-Troitschanskaia, 2011).

Für die Lehrerqualifizierung in den wirtschaftswissenschaftlichen Disziplinen dagegen gibt es trotz ihrer zunehmenden Bedeutung im berufs- und im allgemeinbildenden Bereich[1] nur erste Ansätze, die dort vermittelten Kompetenzen mittels theoretisch und empirisch fundierter Modelle, die valide Aussagen über deren Struktur und Niveaus erlauben, abzubilden (s. Berger et al., 2013; Kuhn, 2014; Zlatkin-Troitschanskaia, Happ et al., 2013). Obgleich unter allen Studiengängen die wirtschaftswissenschaftlichen – mit ca. 15 % der 2.5 Millionen Studierenden an 428 deutschen Hochschulen – am stärksten nachgefragt sind (Statistisches Bundesamt, 2013), lag bis vor Kurzem weder ein geeignetes Kompetenzmodell noch ein dazu passendes deutschsprachiges Messinstrument vor, das den Anforderungen für eine Verwendung im akademischen Feld gerecht wird (Zlatkin-Troitschanskaia, Förster, Brückner, Hansen & Happ, 2013; Tremblay, Lalancette & Roseveare, 2012).

Das seit 2011 vom BMBF geförderte Projekt WiwiKom (s. www.wiwi-kompetenz.de) widmet sich diesem Forschungsdefizit. Es macht sich zur Aufgabe, auf der Grundlage von Testinstrumenten, die im internationalen Raum bereits in großem Maßstab angewandt werden, ein für die deutsche Hochschullandschaft brauchbares modellbasiertes Messinstrument zur Erfassung wirtschaftswissenschaftlicher Fachkompetenz zu entwickeln. Dabei steht die Validierung dieses Tests im Fokus der Analysen, um aus den Ergebnissen der Befragung möglichst tragfähige Schlüsse für den deutschen Hochschulbereich zu begründen.

Das Erreichen dieses umfassenden Ziels wurde methodenintegrativ und zugleich iterativ angestrebt. Im ersten Schritt erfolgte die theoriebasierte Entwicklung eines Kompetenzmodells, das die dimensionale Struktur des Curriculums (nach Inhaltsbereichen wie Marketing, Organisation, Mikro- oder Makroökonomie usw.) sowie die graduellen Ausprägungen der Fachkompetenz angemessen beschreiben kann (z. B. kognitionsbezogene Anforderungen wie Anwenden und Analysieren i. S. v. Lehr-Lernziel-Taxonomien sowie weitere schwierigkeitserzeugende Merkmale). Im nächsten Schritt folgte die Adaptation von zwei internationalen Testinstrumenten an die sprachlichen, kulturellen und inhaltlich-curricularen Gegebenheiten der deutschen wirtschaftswissenschaftlichen Hochschulbildung und die Entwicklung einer deutsch-sprachigen „WiwiKom"-Testversion. Nach umfassenden internationalen Recherchen ist dafür auf den mexikanischen Test „Examen General para el Egreso de la Licenciatura en Administración" (EGEL; Ceneval, 2010) der Mexican Centro Nacional de Evaluación para la Educación Superior (CENEVAL) sowie auf den US-amerikanischen „Test of Understanding in College Economics" (TUCE; Walstad, Watts & Rebeck, 2007) des U. S. Council for Economic Education (CEE) zurückgegriffen worden. Beide Tests sind speziell auf den Hochschulbereich bezogen und werden international bereits in vielen Ländern eingesetzt. Sie umfassen betriebswirtschaftliche (EGEL) sowie volkswirtschaftliche Inhalte (TUCE). Der auf der Basis des Kompetenzmodells und unter Verwendung dieser beiden Instrumente entwickelte deutsche WiwiKom-Test soll es erlauben, die theoretisch postulierte Struk-

1 Neben den Wirtschaftsgymnasien bieten einige Bundesländer (z. B. Hessen, Niedersachsen) das Fach Politik und Wirtschaft auch im allgemeinbildenden Bereich an.

tur und die Niveaus akademischer wirtschaftswissenschaftlicher Fachkompetenz empirisch reliabel und valide abzubilden. Inwieweit dieses Ziel realisiert werden konnte, d. h. inwiefern mit dem WiwiKom-Test die Fachkompetenz von Studierenden und Absolventen wirtschaftswissenschaftlicher Studiengänge in Deutschland zuverlässig und objektiv erfasst werden kann, wurde im Rahmen der methodenintegrativen Validierungsstudien kritisch analysiert (s. Zlatkin-Troitschanskaia, Förster et al., 2013).

Dieser Beitrag geht der Frage nach, inwieweit das mittels des adaptierten und weiterentwickelten Tests erfasste Konstrukt „wirtschaftswissenschaftliche Kompetenz von Studierenden" durch hochschulische und individuelle Einflussfaktoren beeinflusst wird: Nach einer Darstellung des theoretischen Rahmens in Kap. 2 berichtet Kap. 3, wie im Projekt WiwiKom methodenintegrativ vorgegangen wurde, um eine möglichst umfassende Validierung des Tests zu gewährleisten. Im vorliegenden Rahmen können allerdings nur ausgewählte zentrale Ergebnisse dieser methodenintegrativen Studien näher dargestellt werden, wobei der Fokus auf der Analyse inhaltsbezogener Validitätsevidenzen liegt (ausführlicher s. Zlatkin-Troitschanskaia, Förster et al., 2013). Es wird gezeigt, dass das entwickelte WiwiKom-Instrument eine valide Repräsentation der wirtschaftswissenschaftlichen Curricula sowie der tatsächlich gelehrten Inhalte gewährleistet, sodass mit ihm der Erwerb von wirtschaftswissenschaftlicher Fachkompetenz im Studium valide erfasst wird. Die so geschaffene Grundlage ermöglicht u. a. die Untersuchung zweier hochschulpädagogisch zentraler Fragestellungen (Kap. 4), nämlich (a) ob sich die Nutzung der hochschulischen Lehrangebote, wie man erwarten könnte, positiv auf die Ausprägung der wirtschaftswissenschaftlichen Kompetenz der Studierenden auswirkt und (b) welche weiteren kontextuellen Faktoren auf der institutionellen Ebene (z. B. Art der Hochschule) unter Kontrolle individueller Merkmale der Studierenden (z. B. das Vorwissen) deren Fachkompetenz signifikant beeinflussen. Diesen Fragen wird mittels einer Mehrebenenanalyse nachgegangen. Die gefundenen Antworten erfahren in Kap. 5 eine kritische Diskussion.

2. Der theoretische Rahmen

2.1 Zur Modellierung von „wirtschaftswissenschaftlicher Kompetenz"

Betrachtet man die einschlägige internationale Bildungsforschung, so ist festzustellen, dass der Begriff Kompetenz keineswegs einheitlich verstanden wird (vgl. Kuhn & Zlatkin-Troitschanskaia, 2011, sowie die Beiträge in Blömeke, Zlatkin-Troitschanskaia, Kuhn & Fege, 2013). Im Rekurs auf die verbreitet akzeptierte Definition von Weinert (2001) umfasst Kompetenz u. a. die individuell verfügbaren kognitiven Fähigkeiten und Fertigkeiten, die es ermöglichen, Probleme (hier in ökonomischen Situationen bzw. Kontexten) zu lösen. Dabei handelt es sich um *erwerbbare,* also nicht um angeborene Fähigkeiten. Dies bedeutet zugleich, dass sie unter Entwicklungsgesichtspunkten während des Studiums vielfachen Veränderungen, u. a. Vergessens- und Lernprozessen, unterliegen und als deren Resultat betrachtet werden können. Kompetenzen werden

weiterhin als über Einzelsituationen hinweg relativ stabile Dispositionen für professionelles Handeln betrachtet (z. B. verfügbares Fachwissen), die mit volitional-affektiven Komponenten (z. B. Lernmotivation) interagieren. Akademisch erworbene Kompetenzen umschließen somit multiple Facetten und sind daher *mehrdimensional* zu betrachten. Trotz einiger fächerübergreifender Anteile, die in ihnen wirksam werden (wie z. B. forschungsmethodische Fähigkeiten und Fertigkeiten), sind Kompetenzen stets domänenspezifisch konturiert und lassen sich damit von Konstrukten wie allgemeine kognitive Fähigkeiten (z. B. Intelligenz) unterscheiden.

Die aktuelle internationale Kompetenzforschung konzentriert sich bislang überwiegend auf mentale Repräsentationen von Kognitionen als zentrale Dimensionen von Kompetenz (Koeppen, Hartig, Klieme & Leutner, 2008; Sternberg, 2009). Zwar schließen kognitive Dispositionen, theoretisch betrachtet, auch Überzeugungen bzw. Einstellungen ein (z. B. Neuweg, 2011), aber es handelt sich hierbei vorerst um eine analytische Trennung. Empirisch ist in der aktuellen internationalen Forschung ihre separate Erfassung i. S. eines mehrdimensionalen Konstrukts bislang ein Desiderat geblieben. Die Forschungspraxis konzentriert sich vielmehr weitgehend auf (Fach-)Wissen bzw. Wissensstrukturen. Die in der (inter-)nationalen Literatur vorfindlichen Kompetenzkonzepte modellieren sogar nahezu ausschließlich nur das fachwissenschaftliche Wissen und das auf ihm operierende Denken als die Hauptdimensionen des Kompetenzkonstrukts (s. Beck, 1993; Rumelhart & Norman, 1983; Shulman, 1987). Auch Koeppen et al. (2008) nehmen eine solche Eingrenzung des Kompetenzbegriffs vor und verstehen ihn als Bezeichnung einer kontextspezifischen kognitiven Leistungsdisposition.

Den beiden im Projekt WiwiKom adaptierten Tests, EGEL und TUCE, liegt ebenfalls eine fachwissensbezogene Konzeptionierung zugrunde (vgl. Ceneval, 2010; Walstad et al., 2007). Nicht zuletzt aus Gründen der internationalen Anschlussfähigkeit und Vergleichbarkeit wird hier ebenfalls dem kognitiven Verständnis von Fachkompetenz gefolgt, also wirtschaftswissenschaftliches Fachwissen und Denken theoretisch modelliert und empirisch erfasst. Demnach wird das verfügbare wirtschaftswissenschaftliche Fachwissen als kognitive Disposition verstanden, die zur erfolgreichen Bearbeitung wirtschaftlicher Fragestellungen und Situationen erforderlich ist. Zur weiteren Präzisierung des relevanten Fachwissens wird zwischen Inhaltsdimensionen und niveaubezogenen kognitiven Dimensionen differenziert (s. z. B. Alexander, Kulikowich & Schulze, 1994; Beck, 1993). Zur Beschreibung und Charakterisierung der Inhaltsdomäne wird im Projekt WiwiKom einem gängigen fachspezifischen, auch an deutschen Hochschulen etablierten Domänenverständnis gefolgt (s. z. B. Seifried & Ziegler, 2009) und nach den volks- und betriebswirtschaftlichen Subdomänen Mikro- und Makroökonomie, Personal, Finanzierung, Rechnungswesen, Marketing sowie Organisation und Unternehmensführung unterschieden (zur ausführlichen Darstellung und Begründung des theoretischen Wissensmodells s. Zlatkin-Troitschanskaia, Förster et al., 2013).

Bzgl. der niveaubezogenen Dimensionen des wirtschaftswissenschaftlichen Wissens wurden Annahmen zu kognitiven Stufungen entwickelt, die auf möglichst alle seine fachinhaltlichen Dimensionen anwendbar sind. In Anlehnung an Anderson und Krath-

wohl (2001) sowie Walstad et al. (2007) erfolgt eine Differenzierung nach den Stufen (I) Erinnern & Verstehen, (II) Anwenden & Analysieren und (III) Kreieren & Entwickeln, die bezogen auf die Subdomänen weiter spezifiziert ist (z. B. zur Subdomäne Finanzierung s. Förster, Brückner & Zlatkin-Troitschanskaia, im Druck).

Bezugnehmend auf diese zentralen theoretischen Annahmen (ausführlicher s. Zlatkin-Troitschanskaia, Förster et al., 2013) wird im Projekt WiwiKom u. a. die Mehrdimensionalitätshypothese umfassend geprüft. So berichtet der vorliegende Beitrag die Befunde zur *inhaltsbezogenen* Dimensionalitätsanalyse in den beiden Subdomänen der Volkswirtschaftslehre (VWL). Ob die Annahme zweier volkswirtschaftlicher Dimensionen, mikro- und makroökonomisches Wissen, bestätigt werden kann, ist auch Gegenstand der Mehrebenenanalyse, um eine adäquate Wissensscorebildung für die volkswirtschaftlichen Subdimensionen zu ermöglichen (s. Kap. 5).

2.2 Die Beeinflussung des wirtschaftswissenschaftlichen Fachwissens im Mehrebenensystem der Hochschulbildung

Nutzbares Fachwissen lässt sich als zentrales Ergebnis („Outputgröße“) von Bildungsprozessen begreifen, das in Hochschulen vermittelt und (weiter-)entwickelt werden soll. So beschreiben bspw. Helmke und Schrader (2011) mit einem Angebots-Nutzungs-Modell, unter welchen Bedingungen Lernende fachliches Wissen in Lehr-Lernsituationen aufbauen. Hier zeigt sich, dass es insb. die Lehr-Lern-Prozesse, die Anzahl der Lerngelegenheiten sowie die individuellen Voraussetzungen der Lernenden sind, die den Output aus Bildungsprozessen erklären können.

Studien aus der Lehrerbildung für allgemeinbildende Schulen zeigen ebenfalls, dass individuelle Personenmerkmale wie Vorwissen und allgemeine kognitive Fähigkeiten (z. B. Intelligenz) für den (Fach-)Wissenserwerb wesentlich sein können (z. B. Shulman, 1970; Kunter et al., 2011; Blömeke, Bremerich-Voss et al., 2013). Zum möglichen Einfluss der Institution Hochschule gibt es bislang nur ganz wenige Studien, die solche Effekte (z. B. von Art und Anzahl der besuchten Lehrveranstaltungen (LV), des Hochschultyps) in der Studiendomäne Wirtschaftswissenschaften (WiWi) ansatzweise empirisch untersucht haben (Zlatkin-Troitschanskaia, Happ et al., 2013; Happ, Schmidt & Zlatkin-Troitschanskaia, 2013; Kuhn, 2014; s. auch Bouley et al., 2015, in diesem Beiheft).

Aufgrund des defizitären Forschungsstands zu Ursachen bzw. Bedingungsfaktoren des Fachwissenserwerbs und seiner Ausprägung bei Studierenden und Hochschulabsolventen werden in diesem Artikel anhand einer bundesweiten Stichprobe u. a. die Effekte der besuchten LV auf das erfasste Fachwissen in der WiWi-Domäne systematisch geprüft, und zwar unter Kontrolle weiterer struktureller Faktoren der Hochschulinstitution (wie z. B. dem Hochschultyp). Um solche Zusammenhänge beurteilen und unkonfundiert darstellen zu können, werden bedeutsame individuelle Einflussfaktoren wie z. B. das Geschlecht, die Muttersprache, die Abiturnote sowie das fachbezogene Vorwissen kontrolliert, da von diesen ebenfalls ein systematischer Effekt auf das wirt-

schaftswissenschaftliche Fachwissen ausgehen kann (vgl. Zlatkin-Troitschanskaia, Happ et al., 2013).

Bei der Analyse der Einflussfaktoren gilt es weiterhin zu berücksichtigen, dass der Lernraum Hochschule als ein komplexes *Mehrebenensystem* begriffen werden muss, sodass zu prüfen ist, ob sich die Stärke der Einflüsse der verschiedenen strukturellen und individuellen Faktoren zwischen den einzelnen Hochschulinstitutionen bzw. -typen substanziell unterscheidet. Um mögliche Einflüsse seitens der individuellen und der institutionell-kontextuellen Faktoren statistisch differenziert erfassen und betrachten zu können, werden die erwähnten *Mehrebenenanalysen (MLA)* vorgenommen. Dabei sind zwei Ebenen zu modellieren, nämlich die individuelle Ebene 1 der Studierendenmerkmale, die innerhalb der verschiedenen Hochschulinstitutionen auf der kontextuell-institutionellen Ebene 2 geclustert werden.

3. Der methodenintegrative Validierungsprozess im Projekt WiwiKom

Im WiwiKom-Projekt wurden der EGEL (Ceneval, 2010) und der TUCE (Walstad et al., 2007) mithilfe professioneller Translations- und Fachexperten übersetzt und für den deutschen Hochschulsektor umfassend validiert. Da die Darstellung der Ergebnisse zu allen sieben im WiwiKom-Projekt erfassten inhaltlichen Subdimensionen den Umfang des Beitrages sprengen würde, werden hier lediglich die Befunde zur VWL-Fachdomäne dargestellt. Die 60 Aufgaben zu dieser Domäne stammen aus der vierten Edition des TUCE und weisen alle ein geschlossenes Aufgabenformat auf mit je einer richtigen Antwortvorgabe und drei Distraktoren (s. das Beispielitem im Anhang). Innerhalb des Tests wurde von den Testentwicklern eine Differenzierung nach drei Niveaustufen umgesetzt (Erkennen und Verstehen, explizite Anwendung und implizite Anwendung). Je 30 Items des TUCE sind den Subdomänen Mikro- und Makroökonomie zugeordnet. Die Mikroökonomie lässt sich weiter in die Inhaltsbereiche (a) Grundlagen, (b) Märkte und Preise, (c) Unternehmenstheorie, (d) Faktormärkte sowie (e) die mikroökonomische Rolle der Regierung und internationale Mikroökonomie aufteilen. In der Makroökonomie werden folgende Inhaltsbereiche unterschieden: (a) Messung des gesamtwirtschaftlichen Outputs, (b) Angebot und Nachfrage, (c) Geld und Finanzmärkte, (d) Geld- und Fiskalpolitik, (e) politische Debatten und ihre Bewertung sowie (f) internationale Makroökonomie.

Der Validierungsprozess für das Messinstrument wurde methodenintegrativ *(mixed methods)* durchgeführt, wobei verschiedene empirische Evidenzen herangezogen wurden, um die Repräsentativität des Inhalts im Testinstrument und im latenten Konstrukt sicherzustellen. Eine der wesentlichen Analysen zur validen Erfassung von Fachkompetenzen im Hochschulsektor besteht in der Prüfung der angemessenen Abbildung der curricularen Inhalte im Konstrukt sowie im Testinstrument (curriculare Validität). Dem Evidence-centered Design folgend (Mislevy & Haertel, 2006), das dem WiwiKom-Projekt zugrunde gelegt wurde (Brückner, Zlatkin-Troitschanskaia & Förster, 2014), gilt es, zunächst eine umfassende Analyse der Curricula der WiWi-Domäne an den verschiede-

nen Hochschulinstitutionen vorzunehmen. Wird eine Mehrebenenstruktur bei der Validierung berücksichtigt, liegt die zentrale Herausforderung in der standortübergreifenden Überprüfung der Passung des z. B. in Modulhandbüchern ausformulierten sowie des tatsächlich umgesetzten bzw. von den Studierenden absolvierten Curriculums.

Um in einem ersten Schritt die Passung von Konstrukt und Testinstrument hinsichtlich des implementierten Curriculums prüfen zu können (Inhaltsvalidität), war eine Analyse der in Modulhandbüchern und Studienordnungen festgelegten Studieninhalte erforderlich. Hierzu wurden in einer Vorstudie die curricularen Inhalte von 98 Studiengängen an 57 der bundesweit größten wirtschaftswissenschaftlichen Universitäts- und Fachhochschulfakultäten einbezogen (die in den WiwiKom-Erhebungen beteiligten Hochschulen eingeschlossen). Die erfassten Lehr-Lerninhalte wurden in einem Kategorienraster verortet, das auf Basis von Lehrbücheranalysen und Dozenteninterviews erstellt worden war, sodass zum einen die Übereinstimmung zwischen den Testinhalten und den gelehrten Inhalten geprüft werden konnte und zum anderen wirtschaftswissenschaftliche Teil-Curricula ermittelt werden konnten, die an allen oder zumindest an den meisten deutschen Hochschulen gelehrt werden. Hinsichtlich verschiedener LV-Formen (Vorlesung, Seminar etc.) und verpflichtender vs. fakultativer Studieninhalte konnten nicht nur innerhalb einer Hochschulform (z. B. Universität), sondern auch zwischen Hochschulformen (Universitäten und Fachhochschulen (FH)) bedeutsame curriculare Unterschiede festgestellt werden (ausführlicher zu Methoden und Ergebnissen der Inhaltsanalyse in Zlatkin-Troitschanskaia, Förster et al., 2013). Insbesondere hinsichtlich der VWL-Inhalte zeigten sich zwischen den beiden Hochschulformen bedeutsame curriculare Differenzen. An Universtäten ist die Anzahl der ECTS in VWL-Inhaltsgebieten im Durchschnitt fast doppelt so hoch wie an FHs (Lauterbach, 2014). Diese Unterschiede gilt es bei der Analyse der Inferenzen auf Basis der Testwerte zu berücksichtigen. Aus diesem Grund wird die Unterscheidung zwischen FH und Universität in der folgenden MLA berücksichtigt, um zu kontrollieren, inwiefern sich die curricularen Unterschiede in der Ausprägung des Fachwissens wiederfinden lassen.

Im Hinblick auf das WiwiKom-Testinstrument zeigen die Analysen insgesamt, dass die in ihm abgefragten Inhalte relevant und repräsentativ für die jeweilige wirtschaftswissenschaftliche Subdimension sind. Dieser Befund zur curricularen Validität des Tests bestätigt die Ergebnisse weiterer im Projekt WiwiKom durchgeführter Validierungsstudien. So wurden 78 Experten in den (Sub-)Disziplinen der VWL und BWL aus den erfassten Universitäten und FHs gezielt zu spezifischen Aufgabenstellungen interviewt und mithilfe eines standardisierten Online-Fragebogens zur curricularen Repräsentativität und Relevanz der Aufgaben befragt. Zudem wurden diese Expertenbefragungen um u. a. fachdidaktische Analysen der Inhalte zentraler etablierter Lehrbücher (aus Universitäten und FHs) erweitert. Mit dieser Vorgehensweise wurde sowohl die eher statisch-programmatische Perspektive der Lehrbücher und der Modulbeschreibungen als auch eine eher dynamisch-praktizierte Perspektive, nämlich die der Lehrenden mit Blick auf die an ihren Hochschulen tatsächlich gelehrten Inhalte, in die Analysen einbezogen. Die Ergebnisse zeigen, dass die Testaufgaben auch von den Dozenten als repräsentativ für die einzelnen Subdomänen der WiWi und als curricular valide beurteilt wurden.

Insgesamt ergaben diese methodenintegrativen Inhalts- und Curriculaanalysen, dass die in den internationalen Tests thematisierten Inhalte einen Großteil der für den deutschen Hochschulsektor relevanten Fachinhalte abdecken. Dies gilt sowohl für die von uns fokussierten zentralen Subdomänen der BWL (Marketing, Personalwesen, Rechnungswesen, Finanzierung und Organisation/Unternehmensführung) als auch für die beiden Subdomänen Mikro- und Makroökonomie der VWL. Die in den Aufgaben thematisierten Inhalte repräsentieren demnach die Inhalte der jeweiligen Subdomäne.

Neben den curricularen und inhaltlichen Analysen wurden kognitive Interviews mit der Methode des lauten Denkens bei 32 Studierenden eingesetzt, um mentale Prozesse während der Bearbeitung der Testaufgaben (Leighton, Heffernan, Cor, Gokiert & Cui, 2011) zu erfassen und hinsichtlich ihrer Konstruktrelevanz bzw. -irrelevanz zu analysieren.[2] So soll u. a. geprüft werden, ob die zur Lösung der Aufgabe erfassten kognitiven Prozesse der Probanden zentrale Attribute des Konstruktes repräsentieren und nicht durch konstruktirrelevante Prozesse überlagert werden.

Neben der Analyse des in Modulbeschreibungen niedergelegten Curriculums gilt es im Rahmen der mehrschrittigen Validierung das tatsächlich umgesetzte Curriculum zu überprüfen und die resultierenden Testergebnisse mit diesem zu vergleichen. Denn es sollte nicht nur das Konstrukt durch die Inhalte und die Items hinreichend repräsentiert sein; die Beantwortung der Items müsste auch davon abhängen, ob ein bestimmtes Curriculum an einer bestimmten Hochschule absolviert wurde. Bevor diese Prüfung vorgenommen werden konnte, war es zunächst notwendig, den Itemauswahlprozess weiter voranzutreiben. Auf Basis der oben beschriebenen Validierungsanalysen und eines ersten Pretests an zwei Hochschulen im SS 2012 (N = 962) konnten im Ergebnis 204 von insgesamt 403 Items mit geschlossenem Aufgabenformat aus den internationalen Tests übernommen werden; 16 weitere Aufgaben wurden gemeinsam mit Fachexperten für die in den Tests leicht unterrepräsentierten Inhalte entwickelt (ausführlicher zur Itemauswahl und zum Itementwicklungsprozess s. Zlatkin-Troitschanskaia, Förster et al., 2013). Diese 220 Items weisen nach den Curriculumanalysen, den Experteninterviews und der Experten-Onlinebefragung eine gute Passung zwischen dem Konstrukt und dem implementierten Curriculum auf. Es kann also davon ausgegangen werden, dass die Aufgaben des WiwiKom-Tests eine repräsentative Auswahl der für das WiWi-Studium relevanten Inhalte darstellen und keine irrelevanten Inhaltsbereiche abgeprüft werden. Ihre Bearbeitung sollte demnach valide Schlüsse auf die im WiWi-Bereich vermittelten Kompetenzen an deutschen Hochschulen erlauben.

2 Zur Generierung von kognitiven Modellen der Aufgabenbearbeitung auf Basis latenter Konstrukte s. Leighton und Gierl (2007).

4. Empirische Feldstudien: Design und Stichprobe

Da die 220 Aufgaben in allen Validitätsanalyseschritten eine sehr gute Eignung aufwiesen, wurden sie in die erste Feldstudie im WS 2012/13 mit insgesamt 3783 Studierenden an 15 Universitäten und 8 Fachhochschulen aufgenommen. Auf Basis der Datenanalysen mittels klassischer Testtheorie (KTT) und Item-Response-Testtheorie (IRT) erfolgte zunächst eine Itempoolkalibrierung. Dabei wurden einige (in Bezug auf die Reliabilität oder Validität) problematische Items identifiziert und in Kooperation mit Fachexperten modifiziert.

Da im Rahmen der Feldstudie nicht genug Testbearbeitungszeit für die Lösung aller 220 Aufgaben durch jeden Studierenden zur Verfügung stand, wurde ein Booklet-Design eingesetzt. Die 220 Items wurden auf 43 Fragebogenvarianten verteilt. Um eine weitestgehend unverzerrte Schätzung der Item- und Personenparameter zu ermöglichen, griffen wir auf verschiedene Youden-Square Designs zurück (Frey, Hartig & Rupp, 2009).

Im SS 2013 wurde eine zweite Feldstudie an 25 Hochschulen durchgeführt und 3512 Studierende mit dem leicht überarbeiteten WiwiKom-Test befragt. Es wurde eine „paper-and-pencil"-Version eingesetzt. Die Studierenden hatten 45 Minuten Zeit für die Lösung von 30 Fachaufgaben des ihnen vorgelegten Testhefts.

Die im Folgenden referierten Befunde zur Modellierung und Beeinflussung des wirtschaftlichen Fachwissens basieren auf den Daten aus dieser zweiten Feldstudie. Die Teilstichprobe aus der zweiten Feldstudie, die die VWL-Fragen zum TUCE beantwortet hat, umfasst 1150 (bzw. 1151; s. Tab. 1) Studierende aus 25 Hochschulen, davon 18 Universitäten und 7 Fachhochschulen.

	Studierende		**Hochschulen**
	Mikro	**Makro**	
Insgesamt	1.150	1.151	25
Universität	86.6 %	86.7 %	18
Fachhochschule	13.4 %	13.3 %	7
Semester			
Mittelwert	3.64	3.67	–
Median	4	4	–

Tab. 1: Teilstichprobe für die VWL-Dimensionen

5. Ergebnisse

5.1 Dimensionalitätsanalyse und Bildung der Wissensscores

Da die VWL-Fachdimension zwei inhaltliche Subdimensionen umfasst, wird zunächst betrachtet, ob anhand der Daten eine eindimensionale oder eine zweidimensionale Struktur des VWL-Fachwissens mit den beiden Subdimensionen Mikro- und Makroökonomie ermittelt werden kann. Diese Analyse ist neben der Sicherstellung der Konstruktvalidität auch für eine adäquate Wissensscorebildung zur nachfolgenden Mehrebenenanalyse erforderlich. Hierfür wurden zwei Faktorenanalysen durchgeführt. In Tabelle 2 sind die Teststatistiken der konfirmatorischen Faktorenmodelle, auf Basis eines generalisierten linearen Modells mit einem Maximum-Likelihood-Schätzer und einer logit-Funktion als Link-Funktion zwischen den Indikatoren und den Faktoren, abgebildet (s. Rabe-Hesketh, Skrondal & Pickles, 2004).

Die Werte von AIC, BIC und der LR-Chi^2-Test der Differenz der beiden Devianzen machen deutlich, dass das zweidimensionale Modell die Daten signifikant besser beschreiben kann als das eindimensionale, sodass die Analyse zum VWL-Fachwissen im Folgenden getrennt zu jeder Subdimension erfolgen kann, wenngleich die beiden Faktoren mit einer latenten Korrelation von 0.79 einen starken Zusammenhang aufweisen. Auch die Ergebnisse der konfirmatorischen Faktorenanalysen mit Weighted-Least-Square-Schätzern legen dieses Vorgehen nahe.

Die Reliabilität der Subtests zum VWL-Fachwissen (Mikro- & Makroökonomie) wurde im Anschluss an die Dimensionalitätsprüfung durchgeführt. Für die dem Booklet-design angemessenen Reliabilitätsanalysen wurden EAP/PV-Reliabilitäten, die in ihrer Ausprägung mit Cronbach's Alpha vergleichbar sind, und die Spearman-Brown-Approximation verwendet.[3] Dabei konnten für beide Testsubdimensionen zufriedenstellende Testreliabilitäten ermittelt werden (Mikroökonomie: $\alpha = 0.717$; Makroökonomie: $\alpha = 0.788$), die die Reliabilitätskennziffern auf Basis der US-Originaldaten marginal

	Eindimensional	Zweidimensional (Micro + Macro)
Df	120	121
log likelihood	−18,908	−18,897.08
AIC	38,055.99	38,036.15
BIC	38,695.8	38,681.29
LR X^2 Test der Devianzdifferenz (df = 1)		21.84***
Latent correlation		.79

*** p-Value < 0.001

Tab. 2: Ergebnisse der Dimensionalitätsanalyse

3 Die Reliabilität für eine zuverlässige Messung sollte größer als 0.7 sein (Peterson, 1994).

übertreffen (vgl. dazu in den US-Originaldatensätzen: Mikroökonomie: $\alpha = 0.70$; Makroökonomie: $\alpha = 0.77$; Walstad et al., 2007).

Auf der Grundlage der beiden separierbaren Teildimensionen des VWL-Fachwissens und unter Berücksichtigung des gewählten balancierten Incomplete-Block-Designs (Frey et al., 2009), bei dem u. a. die Aufgabenposition zur Kontrolle von Positionseffekten gleichmäßig rotiert wurde, konnten die Personenfähigkeiten latent über Plausible Values (PV) geschätzt werden (Mislevy, 1991). Es wurden je fünf PVs auf Basis eines Rasch-Modells und unter Einbezug der in der MLA verwendeten Kovariaten in einem latenten Regressionsmodell generiert, die jeweils den latenten Wissensscore in Mikro- und Makroökonomie abbilden und so als abhängige Variable in den nachfolgenden MLA genutzt werden. Die Schätzung der PVs basiert auf dem Prinzip der multiplen Imputation (Rubin, 1987) und ermöglicht durch die Berücksichtigung weiterer Kontextvariablen eine weniger konfundierte Schätzung der Personenfähigkeiten auf Populationsebene als andere latente Personenparameter (z. B. weighted maximum likelihood estimates (WLE)).

5.2 Mehrebenenanalysen (MLA)

In der MLA können Einflussgrößen auf das mikro- und makroökonomische Wissen[4] auf zwei Ebenen modelliert werden, wobei die Studierenden auf unterer Ebene, in den Hochschulinstitutionen auf der oberen Ebene geclustert sind. Um den Fragestellungen zum Einfluss der besuchten LV und des Typs der besuchten Hochschulinstitution (FH vs. Universität) auf das VWL-Fachwissen unter Kontrolle weiterer relevanter Kovariaten nachgehen zu können, wurden in diese Modelle verschiedene Kovariate auf der individuellen (Ebene 1) und der kontextuellen Ebene (Ebene 2) integriert und auf das individuelle Fachwissen regressiert. Damit ist es möglich, zwischen den Effekten zur Erklärung von Unterschieden im Fachwissen innerhalb von Institutionen sowie zwischen Institutionen zu differenzieren. Außerdem erlaubt das Verfahren der MLA die Unterscheidung zwischen einer Varianzaufklärung sowohl innerhalb einer Ebene als auch zwischen den Ebenen, sodass eine Beurteilung des Effekts der im Modell betrachteten jeweiligen Kovariaten auf jeder Ebene möglich ist (Raudenbush & Bryk, 2002).

Auf der individuellen Ebene wurden die personenbezogenen Variablen als Kovariate in das Modell aufgenommen. Neben den Variablen, die anzeigen, ob ein Studierender eine bzw. mindestens zwei LV in der jeweiligen VWL-Domäne absolviert hat, werden auf Ebene 1 zusätzlich das Geschlecht, die Muttersprache, die Schulabschlussnote sowie der Abschluss einer kaufmännisch-verwaltenden Ausbildung vor Studienbeginn als Einflussfaktoren berücksichtigt. Bis auf die Schulnote wurden alle Kovariaten

4 Die PVs sind die Indikatoren für das mikro- und makroökonomische Wissen. Die Berechnung der MLA erfolgt separat für jeden einzelnen der fünf PVs. Die Parameter der je fünf Einzelmodelle jeweils für Mikro- und Makroökonomie werden gemäß den Kombinationsregeln nach Rubin (1987) in einem Modell gepoolt.

Individuelle Variablen (Mikro: N = 1150 Studierende; Makro: N = 1151 Studierende)			
		Mikro	**Makro**
Geschlecht	Weiblich	46.4 %	47.7 %
	Männlich	53.6 %	52.3 %
	(Anzahl fehlender Werte)	(3)	(3)
Muttersprache	Deutsch	84.8 %	85.4 %
	Andere	15.2 %	16.6 %
	(Anzahl fehlender Werte)	(2)	(2)
Schulabschlussnote	Mittelwert	2.3	2.3
	(Anzahl fehlender Werte)	(59)	(47)
Berufsausbildung	Ja	83.6 %	80.6 %
	Nein	16.4 %	19.4 %
	(Anzahl fehlender Werte)	(12)	(13)
Lehrveranstaltungen (LV)	Keine LV	33.5 %	51.3 %
	1 LV	49.6 %	34.4 %
	2 oder mehr LV	16.9 %	14.3 %
	(Anzahl fehlender Werte)	(186)	(231)

Tab. 3: Deskriptive Statistiken für die individuellen Variablen

als Dummy-Variablen in die MLA aufgenommen. Die Schulabschlussnote wurde am Stichprobenmittelwert zentriert.[5] In Tabelle 3 sind die deskriptiven Statistiken für die individuellen Variablen dargestellt.

Mittels der MLA kann weiterhin beantwortet werden, in welcher Höhe die Varianzen auf jeder Ebene anfallen und welche Variablen diese Varianzen erklären. Deshalb ist der erste Schritt in der MLA die Schätzung eines Nullmodells (auch Varianzkomponentenmodell genannt) zur Ermittlung (Schätzung) der Varianzen auf beiden Ebenen. Das Nullmodell enthält lediglich eine Konstante mit einem sog. Random-Effekt und keine unabhängigen Variablen (z. B. Raudenbush & Bryk, 2002). Somit findet lediglich eine Aufteilung der Gesamtvarianz auf die beiden Ebenen statt, was bedeutet, dass die Varianz des mikro- bzw. makroökonomischen Wissens in die Varianz innerhalb der Hochschulinstitutionen und in die Varianz zwischen den Institutionen aufgeteilt wird. Durch dieses Modell wird der Interklassenkorrelationskoeffizient (ICC) bestimmt, welcher den Anteil der Gesamtvarianz angibt, der zwischen den Hochschulen anfällt (z. B. Raudenbush & Bryk, 2002). Die Ergebnisse des Nullmodells für mikro- und makroökonomisches Wissen sind in der Tabelle 4 dargestellt.

Wie Tabelle 4 zu entnehmen ist, liegt die erwartete latente Fähigkeit eines zufällig gewählten Studierenden in einer zufällig gewählten Hochschule (der sog. grand mean; z. B. Snijders & Bosker, 2012) für das mikroökonomische Wissen bei −0.258 logits und ist signifikant von 0 verschieden. Für die erwartete latente Fähigkeit des makroökonomischen Wissens wird ein Wert von 0.047 logits vorhergesagt, der in der Grundgesamt-

5 Für eine detaillierte Diskussion zum Umgang mit Kovariaten in der MLA und verschiedener Zentrierungen s. z. B. Enders und Tofighi (2007); Paccagnella (2006).

	Mikro	Makro
Fester Effekt		
Grand mean γ_{00}	−0.258***	0.047
Random effect		
Var(grand mean), u_{0j}	0.07	0.06
Var(Ebene 1), r_{ij}	0.33	0.50
ICC	17.6 %	10.9 %

p-Value: *** < 0.01, ** < 0.05, * < 0.1

Tab. 4: Nullmodelle für das mikroökonomische Wissen und das makroökonomische Wissen

heit nicht signifikant von 0 verschieden ist. Für beide Subdimensionen des VWL-Wissens werden auf Kontextebene Varianzanteile der Gesamtvarianz bestimmt, die nicht vernachlässigt werden sollten (ICC = 17.6 % für das mikro-, ICC = 10.9 % für das makroökonomische Wissen), was die Notwendigkeit und Angemessenheit der mehrebenenanalytischen Betrachtung verdeutlicht. Im folgenden Schritt wird daher ein komplexeres Modell vorgestellt, um die zweite zielleitende Fragestellung beantworten zu können (s. Tab. 5).

Wie in Tabelle 5 zu erkennen ist, haben alle Variablen (bis auf eine) einen signifikanten Einfluss auf das mikro- und das makroökonomische Wissen. Der Intercept von −0.672 logits für die Dimension mikroökonomisches Wissen bzw. −0.596 logits für die Dimension makroökonomisches Wissen stellt den Wert der latenten Fähigkeit für eine weibliche Universitätsstudierende dar, deren Muttersprache nicht Deutsch ist, die keine kaufmännisch-verwaltende Ausbildung absolviert hat, eine mittlere Schulabschlussnote aufweist und die noch keine LV in Mikro- bzw. Makroökonomie besucht hat. Männliche Studierende weisen in beiden Subdimensionen eine höhere Fähigkeit als weibliche Studierende auf.[6] Ebenso haben Studierende mit deutscher Muttersprache einen höheren Score als Studierende mit einer anderen Muttersprache. Studierende mit einer Schulabschlussnote, die unter dem Stichprobenmittelwert liegt, haben einen geringeren Wissensscore als Studierende mit einer Schulabschlussnote, die dem Stichprobenmittelwert (oder besser) entspricht.

Hinsichtlich der zielleitenden Fragestellung kann festgestellt werden, dass die Anzahl der LVs als auch der Typ der Hochschulinstitution die Ausprägung des mikro- bzw. makroökonomischen Wissens signifikant beeinflussen. So fällt der latente Score im mikroökonomischen Wissen im Mittel um etwa 0.169 logits höher aus, wenn Stu-

6 Bereits im Rahmen der Entwicklung des Wirtschaftskundlichen Bildungstests (WBT) als eine deutschsprachige Adaption des amerikanischen Tests of Economic Literacy (TEL) wurde festgestellt, dass männliche Testteilnehmer deutlich höhere Testwerte erzielen (Beck & Wuttke, 2005).

	Mikro	Makro
Fester Effekt		
Intercept	−0.672***	−0.596***
Geschlecht (männlich)	0.211***	0.494***
Muttersprache (Deutsch)	0.254***	0.190**
Kaufmännische Berufsausbildung	0.037	0.219***
Schulabschlussnote (zentriert)	−0.192***	−0.249***
LV in Mikro- bzw. Makroökonomie		
Eine LV	0.169**	0.329***
Mind. zwei LV	0.232***	0.627***
Typ der Hochschulinstitution (FH)	−0.400***	−0.258**
Random effect		
Var(Intercept), u_{0j}	0.04	0.00
Var(Ebene 1), r_{ij}	0.28	0.40
Varianzaufklärung Ebene 1	9.97 %	22.38 %
Varianzaufklärung Ebene 2	60.00 %	100.00 %

p-Value: *** < 0.01, ** < 0.05, * < 0.1

Tab. 5: Random-Intercept-Modell für das mikroökonomische und das makroökonomische Wissen mit Individualvariablen als Kovariaten

dierende eine mikroökonomische LV besucht haben, im Vergleich zu Studierenden, die noch keine einschlägige LV besucht haben. Studierende, die mindestens zwei oder mehr LV zur Mikroökonomie besucht haben, weisen im Mittel einen etwa 0.232 logits höheren Score auf als Studierende ohne mikroökonomische LV. Die Effekte der LV sind im makroökonomischen Wissensscore noch stärker ausgeprägt. Hier weisen Studierende, die eine makroökonomische LV besucht haben, im Mittel einen um 0.329 logits höheren Score auf und Studierende, die mindestens zwei makroökonomische LV absolvierten, haben im Mittel sogar einen um 0.627 logits höheren Score. Unter Kontrolle der Anzahl der besuchten LV sowie der weiteren individuellen Variablen nimmt auf der institutionellen Ebene der Typ der Hochschulinstitution einen signifikanten Einfluss auf den latenten Score in beiden Dimensionen. So weisen Studierende der FH bzgl. des mikroökonomischen Wissens im Mittel einen um etwa 0.4 logits und bzgl. des makroökonomischen Wissens einen um 0.258 logits geringeren Score auf als Studierende an Universitäten.

Die einbezogenen Kovariaten klären auf Individualebene für das mikroökonomische Wissen 9.97 % der Residualvarianz und für das makroökonomische Wissen 22.38 % der Residualvarianz auf. Die Varianzaufklärung auf institutioneller Ebene ist mit 60 % bzgl. des mikroökonomischen Wissens und mit 100 % bzgl. des makroökonomischen

Wissens sehr hoch. Zurückzuführen ist dieser Befund nicht allein auf die Aufnahme der kontextuellen Kovariaten in das Modell, sondern auch auf die individuumbezogenen Variablen, die nicht nur innerhalb der Hochschulen variieren, sondern auch zwischen den Hochschulen systematische Variationen aufweisen. So gibt es bspw. Hochschulen, deren Studierende im Mittelwert eine geringere Schulabschlussnote aufweisen als diejenigen anderer Hochschulen.

6. Diskussion und Fazit

Die empirischen Befunde zeigen, dass alle im Modell betrachteten individuellen Merkmale einen signifikanten Effekt auf das makroökonomische Wissen und, bis auf die kaufmännische Ausbildung, auch auf das mikroökonomische Wissen aufweisen. Zudem konnte auch mit dem Hochschultyp ein Faktor ermittelt werden, der auf kontextueller, hochschulisch-struktureller Ebene Einfluss auf das Fachwissen der Studierenden nimmt. Insgesamt deuten die Ergebnisse aus der MLA darauf hin, dass der theoretisch postulierte Effekt des Lernorts Hochschule auf das VWL-Fachwissen (in Gestalt der besuchten LV) tatsächlich nachgewiesen werden kann und sich das Fachwissen der Studierenden in Abhängigkeit von der besuchten Hochschule unterscheidet. Der Anteil der Varianz, der auf Ebene der Hochschule generiert wird, liegt bei 11–18 %. Dies deutet darauf hin, dass die Wahl der Hochschule für den Fachkompetenzerwerb von Relevanz ist. In der Studie schneiden die Studierenden der Universitäten besser ab als die Studierenden der FH. Dies könnte als Indiz dafür gedeutet werden, dass die Universitäten in der Vermittlung der hier betrachteten kognitiven Kompetenzdimensionen höhere Ausprägungen erzielen. Allerdings konnten wir bei den Curriculumanalysen u. a. feststellen, dass an den FHs im Durchschnitt deutlich weniger VWL-Inhalte gelehrt werden. Daher sollte dieser Befund unter Kontrolle weiterer hochschulinstitutioneller Merkmale noch genauer untersucht werden.

Neben den Effekten auf der zweiten Ebene ist hervorzuheben, dass auch hochschulbezogene Einflüsse auf der ersten Ebene zu verzeichnen sind. Das hier aufgenommene Merkmal der besuchten VWL-Kurse ist eine hochschulbezogene Variable auf Individualebene, die den Effekt der besuchten Hochschule noch erhöht. Die besuchten LV erbringen eine beachtliche Erklärungsleistung in beiden VWL-Dimensionen.

Ungeachtet der Varianzaufklärung auf Hochschulebene bleibt jedoch zu konstatieren, dass die Individualebene für die Erklärung des VWL-Fachwissens einen deutlich höheren Beitrag leistet. Die MLA-Modelle zeigen, dass die individuellen Voraussetzungen, die die Studierenden mitbringen, für die Testergebnisse, d. h. für die Ausprägungen der entsprechenden Fachwissensdimensionen, wesentlich bedeutsamer sind als institutionelle und curriculare Unterschiede zwischen den Hochschulen. So bilden die Merkmale Geschlecht, Muttersprache, Schulabschlussnote und z. T. auch das Vorwissen in Form einer absolvierten Berufsausbildung entscheidende Erklärungsfaktoren. Demnach haben Studierende, die z. B. eine kaufmännisch-verwaltende Berufsausbildung absolviert haben, Vorteile gegenüber Studierenden ohne eine solche Ausbildung. Diese

Effekte werden in den beiden betrachteten Testdimensionen (Makro- und Mikroökonomie) auch bei zunehmender Studiendauer kaum geringer. Die Befunde deuten demnach darauf hin, dass es im Hochschulbereich (noch) nicht gelingt, Eingangsunterschiede im Fachwissen – wie zwischen den Geschlechtern oder verschiedenen Muttersprachler/innen – im Studium abzubauen. Hier gilt es weiter zu erforschen, ob die gefundenen Effekte z. B. auf bestimmte Testformen (Beck & Wuttke, 2005), auf das Lernverhalten oder auf geschlechtsspezifische Verarbeitungs- und Lernprozesse der Studierenden zurückgeführt werden können.

Kritisch hervorzuheben ist, dass die hier betrachteten Kovariaten der Ebene 2, soweit sie in unsere Studie aufgenommen werden konnten, unspezifisch sind hinsichtlich der Erklärung von mikro- vs. makroökonomischem Fachwissen. Um zu untersuchen, ob bspw. Lehrqualitätsmerkmale das Fachwissen der Studierenden beeinflussen, sollten daher in den nächsten Forschungsschritten zusätzliche Kontextmerkmale in die Analyse aufgenommen werden. Die Prüfung von Kausalitätshypothesen schließlich bedarf zudem eines Längsschnittdesigns, mit dem die Entwicklung des Fachwissens während des Studiums und seine Prädiktoren erfasst werden können.

Anhang

Die beiden einzigen Cola-Hersteller eines Landes (A-Cola und B-Cola) entscheiden über Preiserhöhungen und -senkungen für ihre Colas. Die nachfolgende Tabelle zeigt die Preisstrategien der Unternehmen und den zu erwartenden Gewinn bzw. Verlust beider Unternehmen in Millionen Euro.

		Preisstrategie A-Cola	
		Hoher Preis	**Niedriger Preis**
Preisstrategie B-Cola	**Hoher Preis**	A-Cola + 100 B-Cola + 100	A-Cola + 250 B-Cola − 50
	Niedriger Preis	A-Cola − 50 B-Cola + 250	A-Cola + 50 B-Cola + 50

Wenn beide Unternehmen davon ausgehen, dass die Mehrzahl der Verbraucher bald keine Cola mehr trinken, sondern auf andere Produkte umsteigen wird, was ist die logische Folge?

- ☐ A-Cola und B-Cola wählen beide einen niedrigen Preis.
- ☐ A-Cola und B-Cola wählen beide einen hohen Preis.
- ☐ A-Cola wählt einen niedrigen Preis, B-Cola wählt einen hohen Preis.
- ☐ A-Cola wählt einen hohen Preis, B-Cola wählt einen niedrigen Preis.

Beispielitem

Literatur

Alexander, P. A., Kulikowich, J. M., & Schulze, K. S. (1994). How Subject-Matter Knowledge Affects Recall and Interest. *American Educational Research Journal, 31*(2), 313–337.

Anderson, L. W., & Krathwohl, D. R. (Hrsg.) (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Boston: Allyn & Bacon.

Beck, K. (1993). *Dimensionen der ökonomischen Bildung. Messinstrumente und Befunde*. Nürnberg: Universität Erlangen-Nürnberg (Abschlussbericht zum DFG-Projekt Az. II A 4-Be 1077/3).

Beck, K., & Wuttke, E. (2005). Ökonomische Kompetenz. In D. Fey, L. von Rosenstiel & C. G. Hoyos (Hrsg.), *Wirtschaftspsychologie* (S. 279–283). Weinheim/Basel: Beltz.

Berger, S., Fritsch, S., Seifried, J., Bouley, F., Mindnich, A., Wuttke, E., Schnick-Vollmer, K., & Schmitz, B. (2013). Entwicklung eines Testinstruments zur Erfassung des fachlichen und fachdidaktischen Wissens von Studierenden der Wirtschaftspädagogik – Erste Erfahrungen und Befunde. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften* (Lehrerbildung auf dem Prüfstand, Sonderheft, S. 93–107). Landau: Verlag Empirische Pädagogik.

Blömeke, S., Bremerich-Voss, A., Kaiser, G., Nold, G., Haudeck, H., Keßler, J.-U., & Schwippert, K. (Hrsg.) (2013). *Professionelle Kompetenzen im Studienverlauf: Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrerausbildung aus TEDS-LT*. Münster: Waxmann.

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., & Fege, J. (Hrsg.) (2013). *Modeling and Measuring Competencies in Higher Education*. Rotterdam: Sense Publishers.

Bouley, F., Berger, S., Fritsch, S., Wuttke, E., Seifried, J., Schnick-Vollmer, K., & Schmitz, B. (2015). Der Einfluss von universitären und außeruniversitären Lerngelegenheiten auf das Fachwissen und fachdidaktische Wissen von angehenden Lehrkräften an kaufmännisch-berufsbildenden Schulen. *Zeitschrift für Pädagogik, 61. Beiheft*, 100–115.

Brückner, S., Zlatkin-Troitschanskaia, O., & Förster, M. (2014). Relevance of Adaptation and Validation for International Comparative Research on Competencies in Higher Education – A Methodological Overview and Example from an International Comparative. Project within the KoKoHs Research Program. In F. Musekamp & G. Spöttl (Hrsg.) (im Druck), *Competence in Higher Education and the Working Environment. National and International Approaches for Assessing Engineering Competence*. Frankfurt a. M.: Peter Lang (Vocational Education and Training: Research and Practice).

Ceneval (2010). *Examen General para el Egreso de la Licenciatura en Administración* (EGEL-ADMON). Mexico City: Centro Nacional de Evaluación para la Educación Superior.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138.

Förster, M., Brückner, S., & Zlatkin-Troitschanskaia, O. (im Druck). Assessing financial knowledge of university students in Germany. In O. Zlatkin-Troitschanskaia & R. J. Shavelson (Hrsg.), *Assessment of Domain-specific Professional Competencies in Empirical Research in Vocational Education and Training* (Special Issue).

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53.

Happ, R., Schmidt, S., & Zlatkin-Troitschanskaia, O. (2013). Der Stand des wirtschaftswissenschaftlichen Fachwissens von Bachelorabsolventen der Universität und der Fachhochschule. In U. Faßhauer, B. Fürstenau & E. Wuttke (Hrsg.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung 2013* (S. 73–85). Opladen: Budrich.

Helmke, A., & Schrader, F.-W. (2011). Vom Angebots-Nutzungs-Modell zur Unterrichtsentwicklung. In A. Bartz, H.-J. Brandes & S. Engelke (Hrsg.), *Praxishilfen für die mittlere Führungsebene in der Schule* (S. 3–6). Köln: Link.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Journal of Psychology, 216*(2), 61–73.

Kuhn, C. (2014). *Fachdidaktisches Wissen von Lehrkräften im kaufmännisch-verwaltenden Bereich. Modellbasierte Testentwicklung und Validierung* (Empirische Berufsbildungs- und Hochschulforschung, Bd. 2). Landau: Verlag Empirische Pädagogik.

Kuhn, C., & Zlatkin-Troitschanskaia, O. (2011). *Assessment of Competencies among University Students and Graduates – Analyzing the State of Research and Perspectives* (Working Paper: Business Education, 59). Mainz: Johannes Gutenberg-Universität.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Lauterbach, O. (2014). *Erfassung wirtschaftswissenschaftlicher Fachkompetenz von Studierenden unter Berücksichtigung des Hochschultyps*. Präsentation auf dem 24. Kongress der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE) am 10. 03. 2014 in Berlin.

Leighton, J. P., & Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education. Theory and Applications*. Cambridge: Cambridge University Press.

Leighton, J. P., Heffernan, C., Cor, M. K., Gokiert, R. J., & Cui, Y. (2011). An Experimental Test of Student Verbal Reports and Teacher Evaluations as a Source of Validity Evidence for Test Development. *Applied Measurement in Education, 24*(4), 324–348.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J., & Haertel, G. (2006). *Implications of Evidence-Centered Design for Educational Testing* (Draft PADI Technical Report 17). Menlo Park: SRI International.

Neuweg, G. H. (2011). Das Wissen der Wissensvermittler. Problemstellungen, Befunde und Perspektiven der Forschung zum Lehrerwissen. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 451–477). Münster: Waxmann.

Paccagnella, O. (2006). Centering or Not Centering in Multilevel Models? The Role of the Group Mean and the Assessment of Group Effects. *Evaluation Review, 30*(1), 66–85.

Peterson, R. A. (1994). A Meta-Analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research, 21*(2), 381–391.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167–190.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (Bd. 1, 2. Aufl.). Thousand Oaks: Sage.

Rubin, D. B. (1987). *Multiple imputations for nonresponse in surveys*. New York: John Wiley and Sons.

Rumelhart, D. E., & Norman, D. A. (1983). *Representation in memory*. San Diego: University of California.

Seifried, J., & Ziegler, B. (2009). Domänenbezogene Professionalität. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.). *Lehrprofessionalität* (S. 83–92). Weinheim/Basel: Beltz.

Shulman, L. S. (1970). Cognitive Learning and the Educational Process. *The Journal of Medical Education, 45*(11), 90–100.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *The Elementary School, 57*(1), 1–22.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2. Aufl.). London: Sage.

Statistisches Bundesamt (2013). *Bildung und Kultur – Studierende an Hochschulen*. Wiesbaden: Statistisches Bundesamt. https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Hochschulen/StudierendeHochschulenEndg2110410137004.pdf?__blob=publicationFile [24. 05. 2014].

Sternberg, R. (2009). *Cognitive Psychology*. Wadsworth: Cengage Learning.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of Higher Education Learning Outcomes – Feasibility Study Report. Volume 1 – Design and Implementation*. Paris: OECD.

Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of understanding in college economics: Examiner's manual* (4. Aufl.). New York: National Council on Economic Education.

Weinert, F. E. (2001). Competencies and Key Competencies: Educational Perspective. In N. J. Smelser & P. B. Baltes (Hrsg.), *International Encyclopedia of the Social and Behavioral Sciences* (4. Aufl., S. 2433–2436). Amsterdam: Elsevier.

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., Hansen, M., & Happ, R. (2013). Modellierung und Erfassung der wirtschaftswissenschaftlichen Fachkompetenz bei Studierenden im deutschen Hochschulbereich. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften* (Lehrerbildung auf dem Prüfstand, Sonderheft, S. 108–133). Landau: Verlag Empirische Pädagogik.

Zlatkin-Troitschanskaia, O., Happ, R., Förster, M., Preuße, D., Schmidt, S., & Kuhn, C. (2013). Analyse der Ausprägung und Entwicklung der Fachkompetenz von Studierenden der Wirtschaftswissenschaften und der Wirtschaftspädagogik. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.), *Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften* (Lehrerbildung auf dem Prüfstand, Sonderheft, S. 69–92). Landau: Verlag Empirische Pädagogik.

Abstract: Research on the influence of learning opportunities in business and economics in higher education on the development of subject-specific competency over the course of studies is scarce. In our study we analyze data of the WiwiKom project, which were gathered at 33 higher education institutions, to identify the ways in which students' competency in economics is influenced by individual and contextual factors. A multilevel analysis is conducted in which the hierarchical structure of the higher education system is taken adequately into consideration, thus allowing a non-confounded estimation of the influences of the learning opportunities perceived in higher education. Results indicate significant and inter-individual differences and are discussed comprehensively.

Keywords: Subject-Specific Competency, Economics, Mixed Methods, Multilevel Analysis, Validity

Anschrift der Autor(inn)en

Prof. Dr. Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-Universität Mainz, Fachbereich Rechts- und Wirtschaftswissenschaften, Jakob Welder-Weg 9, 55099 Mainz, Deutschland
E-Mail: lstroitschanskaia@uni-mainz.de

Jun.-Prof. Dr. Manuel Förster, Johannes Gutenberg-Universität Mainz, Fachbereich Rechts- und Wirtschaftswissenschaften, Jakob Welder-Weg 9, 55099 Mainz, Deutschland
E-Mail: manuel.foerster@uni-mainz.de

Dipl.-Hdl. Susanne Schmidt, Johannes Gutenberg-Universität Mainz, Fachbereich Rechts- und Wirtschaftswissenschaften, Jakob Welder-Weg 9, 55099 Mainz, Deutschland
E-Mail: susanne.schmidt@uni-mainz.de

Dipl.-Hdl. Sebastian Brückner, Johannes Gutenberg-Universität Mainz, Fachbereich Rechts- und Wirtschaftswissenschaften, Jakob Welder-Weg 9, 55099 Mainz, Deutschland
E-Mail: brueckner@uni-mainz.de

Univ.-Prof. i. R. Dr. Klaus Beck, Johannes Gutenberg-Universität Mainz, Fachbereich Rechts- und Wirtschaftswissenschaften, Jakob Welder-Weg 9, 55099 Mainz, Deutschland
E-Mail: beck@uni-mainz.de

Gabriele Kaiser

Erfassung berufsbezogener Kompetenzen von Studierenden

Ein Kommentar

Die Kompetenzentwicklung von Studierenden im Hochschulsektor war bisher kaum im Fokus von standardisierten Messungen, im Gegensatz zur Kompetenzentwicklung von Lernenden im Schulbereich, die seit mehr als einem Jahrzehnt regelhaft erhoben wird. Im Hochschulsektor ist die Frage nach der Effektivität der universitären Ausbildung zuerst im Bereich der Lehrerbildung nachdrücklich gestellt worden, auch weil die Ausbildung von Lehrkräften in vielen Ländern Kritik ausgesetzt war und verändert wurde, ohne empirische Evaluation ihrer Wirkungen. Von daher kann die Forschung zur Effektivität der Lehrerausbildung als Vorreiter für die Diskussionen angesehen werden, die mit dem BMBF-Förderprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor" an deutschen Hochschulen auf breiter Ebene angestoßen wurden. Der Kommentar orientiert sich daher aufgrund dieser Vorreiterrolle als Interpretationshintergrund an Arbeiten zur empirischen Evaluation der Effektivität der Lehrerausbildung.

Die empirische Evaluation der Lehrerbildung wurde 2008 entscheidend vorangetrieben durch eine Large-Scale-Studie der International Association for the Evaluation of Educational Achievement (IEA) zur Effektivität der Ausbildung von Mathematiklehrkräften für die Primarstufe und die Sekundarstufe, die Studie „Teacher Education and Development Study: Learning to Teach Mathematics" (TEDS-M). TEDS-M entwickelte unter Bezug auf den Weinert'schen Kompetenzbegriff einen theoretischen Rahmen zur Beschreibung von Lehrerprofessionswissen. Dabei orientiert sich die Beschreibung der von zukünftigen Lehrkräften zu erwerbenden Kompetenzen unter Bezug auf Weinert (1999) auf die beruflichen Anforderungssituationen, mit denen die zukünftigen Lehrkräfte in ihrer späteren Berufspraxis konfrontiert werden (z. B. neben TEDS-M auch bei COACTIV, siehe Baumert & Kunter, 2006). Diese allgemeine Kompetenzdefinition ist damit auf breiter Ebene anschlussfähig, auch für die Beschreibung der Kompetenzentwicklung von Studierenden im Hochschulsektor, die nicht Lehramtsstudierende sind. Dabei wird in der einschlägigen Diskussion um Lehrerprofessionswissen zwischen professionellem Wissen und affektiv-motivationalen Charakteristika unterschieden. In Anschluss an Arbeiten von Shulman (1986, 1987) und Bromme (1992) wird professionelle Kompetenz bestehend aus professionellem Wissen und affektiv-motivationalen Charakteristika beschrieben (siehe den Überblicksartikel von Blömeke & Delaney, 2012).

In den wegweisenden Arbeiten von Shulman (1986, 1987) wird professionelles Wissen konzeptualisiert mit den drei Domänen: Fachwissen, Fachdidaktisches Wissen, Pädagogisches Wissen, wobei fachdidaktisches Wissen curriculares Wissen mit einschließt.

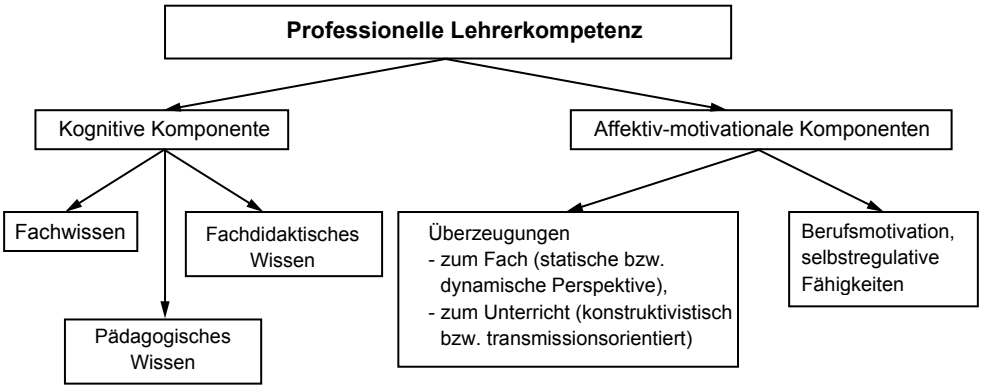


Abb. 1: Modell professioneller Lehrerkompetenz (aus Blömeke, Suhl & Döhrmann, 2012, S. 423)

Insbesondere die Konzeptualisierung von fachdidaktischem Wissen durch Shulman (1987) als „that special amalgam of content and pedagogy that is uniquely the province of teachers, their own special form of professional understanding" (S. 8) hat die Diskussion vorangetrieben und entscheidend beeinflusst. Shulman (1987) selber ist wohl davon ausgegangen, dass fachdidaktisches Wissen eine eigenständige Domäne darstellt, indem er *pedagogical content knowledge* weiter konzeptualisiert als: „It represents the blending of content and pedagogy into an understanding of how particular topics, problems, or issues are organized, represented, and adapted to the diverse interests and abilities of learners, and presented for instruction" (S. 8). Affektiv-motivationale Charakteristika wie *beliefs* zum Lehren und Lernen oder zur Struktur des Unterrichtsfachs und Berufsmotivation sowie Selbstregulation werden ebenfalls als unverzichtbare Bestandteile von professioneller Kompetenz von Lehrkräften angesehen (siehe Abb. 1).

TEDS-M zeigt auf, dass viele Mathematiklehrkräfte am Ende ihrer Ausbildung nach dem Referendariat über ein mathematisches und mathematikdidaktisches Wissen verfügen, das nicht den Anforderungen eines qualitätsvollen Mathematikunterrichts genügt (für Details siehe Blömeke, Kaiser & Lehmann, 2010a, 2010b). Die Übertragung des theoretischen Rahmens von TEDS-M auf die gering strukturierten Domänen der Ausbildung für Deutsch- und Englischlehrkräfte in der Studie TEDS-Learning to Teach (TEDS-LT) macht zentrale Defizite auch in diesen Lehramtsstudiengängen deutlich. Dabei zeigt sich eine interessante Strukturgleichheit der Ergebnisse über die drei Unterrichtsfächer Deutsch, Englisch, Mathematik. Angesichts der Unterschiedlichkeit dieser drei Fächer deutet diese Gleichheit darauf hin, dass trotz aller Fachspezifität der Lehramtsausbildung gewisse Gemeinsamkeiten über die zentralen Kernfächer hinweg existieren. Danach ist im Lehramtsstudium eher nicht von einem kumulativen Lernen auszugehen, wobei der universitäre Standort für den Kompetenzerwerb hochbedeutsam ist (für Details siehe Blömeke et al., 2013, 2011).

Zentrales Ziel neuerer Arbeiten zur Lehrerprofessionalisierung ist die Frage, durch welche Struktur handlungsbezogenes Lehrerprofessionswissen beschrieben werden

kann, welche Facetten bedeutsam sind und wie diese einander beeinflussen. Die Beschreibung der von zukünftigen Lehrkräften zu erwerbenden Kompetenzen auf die beruflichen Anforderungssituationen, mit denen die zukünftigen Lehrkräfte in ihrer späteren Berufspraxis konfrontiert werden, führt in neueren Arbeiten zur Kompetenzforschung im Bereich Lehrerbildung zu einer erweiterten Auffassung der Transformation von Kompetenz in Performanz, in der neben kognitiven Wissenskomponenten situations- und verhaltensnahe Fähigkeiten eine Rolle spielen. So unterscheiden Blömeke, Gustafsson und Shavelson (im Druck) neben der kognitiven und der affektiven Komponente situationsspezifische Fähigkeiten, nämlich Wahrnehmen, Interpretieren und Entscheiden, die dann zu einem beobachtbaren Verhalten führen. Aktuelle deutschsprachige Studien – z. B. die Follow-up-Studie von TEDS-M – unterscheiden als Facetten von Lehrerexpertise die präzise Wahrnehmung von Unterrichtssituationen, deren zielangemessene Analyse und Interpretation sowie die flexible Reaktion darauf (für Details siehe Blömeke et al., 2014).

Die in diesem Beiheft erscheinenden Beiträge eröffnen nun andere, bisher empirisch nicht erforschte Bereiche der berufsbezogenen Kompetenzmessung.

Der Beitrag von Hammer et al. zur „Kompetenz von Lehramtsstudierenden in Deutsch als Zweitsprache: Validierung des GSL-Testinstruments“ setzt bei einem zentralen Thema der aktuellen Bildungsdiskussion an, nämlich einer zunehmend mehrsprachigen Schülerschaft im deutschen Schulwesen und der Frage der für einen angemessenen Unterricht nötigen Kompetenzen von Lehramtsstudierenden im Bereich Deutsch als Zweitsprache. Das Projekt DaZKom orientiert sich dabei an den theoretischen Rahmenkonzeptionen von TEDS-M und der Vorgängerstudie MT21 und wählt daher das Schulfach Mathematik als Bezugsdisziplin. Gleichzeitig wird davon ausgegangen, dass die Verbindung von fachdidaktischer Professionalität mit Kompetenzen im Bereich Deutsch als Zweitsprache als generische Kompetenz angesehen werden kann. So wird einerseits ein sprachlich orientiertes Kompetenzmodell bestehend aus Fachregister (Sprache), Mehrsprachigkeit (Lernprozess), Didaktik (Lehrprozesse) entwickelt, gleichzeitig werden aber unter Bezug auf die oben dargestellten Ansätze von Shulman (1986, 1987) Lehrerkompetenzen beschrieben als Fachwissen (gemessen als linguistisches Wissen), fachdidaktisches Wissen (gemessen als mathematikdidaktisches Wissen) und pädagogisches Wissen.

Dieser Ansatz spiegelt das grundsätzliche Problem der Diskussion um die Vermittlung von Kompetenzen zu Deutsch als Zweitsprache wider; diese Kompetenzen werden einerseits generisch ohne jeden Domänenbezug konzeptualisiert, sind aber andererseits stark sprachlich ausgerichtet und können ohne Fachbezug nur schwerlich unterrichtlich wirksam werden. Die unterschiedlichen Modelle der Implementierung von DaZ an Universitäten spiegeln dieses Dilemma wieder, bringen aber konzeptuelle Probleme. So bleibt unklar, inwieweit wirklich generische Kompetenzen erhoben wurden, insbesondere ist fraglich, ob hier wirklich fachdidaktisches Wissen im Sinne von Shulmans Amalgam-Hypothese erhoben wurde. Hätte dann nicht das mathematikdidaktische Wissen auf sprachliche Aspekte reduziert werden müssen ohne Rückgriff auf TEDS-M-Instrumente? Diese Vermutung wird durch das Ergebnis der Studie gestützt, dass Studie-

rende mit hohem mathematikdidaktischem Wissen über keine höhere Sensibilität für mathematikspezifische Fachsprachen verfügen. Das Ergebnis zeigt auch, dass das Wissen über sprachlich bedingte Schwierigkeiten von mehrsprachigen Lernenden im Mathematikunterricht noch nicht in der Lehramtsausbildung angekommen ist.

Der Beitrag von Riese et al. zur „Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik" knüpft ebenfalls an die Shulman'sche Beschreibung von professionellem Wissen an. Allerdings erfolgt dann eine Fokussierung auf „physikbezogenes Professionswissen" mit einem Ausblenden der Domäne des pädagogischen Wissens. Unter Bezug auf in der Physikdidaktik akzeptierte Ansätze wird eine eigene Wissensdomäne eingefügt, das „Erklärungswissen", welches Lehrpersonen für eine sachlich angemessene und schülergemäße Kommunikation naturwissenschaftlicher Sachverhalte benötigen. Diese Wissensdomäne weist zwar starke Bezüge zum fachdidaktischen und fachlichen Wissen auf, wird aber als eigenständiger Wissensbereich aufgefasst. Das Projekt geht bei der Erhebung des Lehrerprofessionswissens innovative Wege dahingehend, dass nicht nur schriftliche Leistungstests eingesetzt werden, sondern auch ein qualitatives videobasiertes Instrument für die Messung des Erklärungswissens. Mittels Einsatz von praxisnahen Rollenspielen, die videografiert werden, soll eine bessere Abbildung realen Lehrerhandelns erreicht werden. Damit schließt die Studie an oben skizzierte neuere Ansätze zu Lehrerkompetenzen an, die stärker unterrichtsnahe Fähigkeiten in Lehrerkompetenzen integrieren.

Es drängt sich als offene Frage auf, inwieweit die im Projekt vorgenommene Einschränkung auf die Mechanik, die zwar ein zentraler, aber auch relativ überschaubarer Bereich der universitären Physik darstellt, die Ergebnisse beeinflusst und einschränkt. Die Übertragbarkeit der Ergebnisse auf andere physikalische Themengebiete scheint in Folgestudien geboten. Die Autor(inn)en weisen auf einige Einschränkungen der Konzeptualisierung hin, die in der weiteren Datenauswertung bzw. in Folgeprojekten geklärt werden sollen. In der Mathematikdidaktik, die ja traditionell weite Überschneidungen mit der Physikdidaktik aufweist, sind im Rahmen von Ergänzungsstudien zu TEDS-M einige Konzeptualisierungen entwickelt worden, die Anregungspotenzial für das Projekt ProfiLe_P haben können. So weisen Buchholtz, Kaiser und Blömeke (2014) in ihren Analysen der Studie TEDS-LT auf die Mehrperspektivität mathematikdidaktischen Wissens hin, das sowohl von der Mathematik, der Psychologie, der Erziehungswissenschaft und der Allgemeinen Didaktik geprägt ist. Sie schlagen daher eine Unterteilung des mathematikdidaktischen Wissens in die Subdimension „stoffbezogenes mathematikdidaktisches Wissen" vor, worunter stofflich geprägte Fragestellungen des Lehrens und Lernens von Mathematik sowie eine fachlich geprägte Diagnostik von Schülerlösungen verstanden werden. In der Subdimension „unterrichtsbezogenes mathematikdidaktisches Wissen" werden erziehungswissenschaftlich-psychologische Perspektiven berücksichtigt wie Konzepte mathematischer Bildung, Curricula und Bildungsstandards, Leistungsdiagnostik und Ähnliches. Eine ähnliche Unterteilung könnte auch für das physikdidaktische Wissen ertragreich sein.

Für die im Projekt vorgenommene Trennung von vertieftem Schulwissen und universitärem Wissen kann zur begrifflichen Klärung der in der Mathematikdidaktik be-

deutsame Ansatz von Felix Klein zur „Elementarmathematik vom höheren Standpunkte aus“ hilfreich sein. In dem Projekt TEDS-Telekom, in dem im Anschluss an die TEDS-M-Studie längsschnittlich die Entwicklung des Lehrerprofessionswissens in den ersten drei Studienjahren erhoben wurde, wurde das mathematische Wissen in universitäre Mathematik und Elementarmathematik vom höheren Standpunkt unterteilt, da insbesondere Letzterer in einschlägigen mathematikdidaktischen Ansätzen eine Vermittlungsfunktion von Schulmathematik und Universitätsmathematik zugesprochen wird (Buchholtz & Kaiser, 2013). Auch die Berücksichtigung des erziehungswissenschaftlichen Wissens erscheint im Lichte der Erkenntnisse von TEDS-M, COACTIV-R und BILWISS unverzichtbar, da es zwischen diesen drei Domänen zentrale Wechselwirkungen gibt (König & Seifert, 2012; Kunina-Habenicht et al., 2013; Kunter et al., 2011).

Der Beitrag von Dunekacke et al. zum Thema „Mathematikdidaktische Kompetenz von Erzieherinnen und Erziehern: Validierung des KomMa-Leistungstests durch die videogestützte Erhebung von Performanz“ nimmt einerseits eine bisher völlig vernachlässigte Berufsgruppe in den Fokus, nämlich Erzieherinnen und Erzieher. Die Studie untersucht das professionelle Wissen von angehenden Erzieher(innen) und bezieht sich dabei einerseits auf Ansätze aus der Lehrerprofessionsforschung und den dort entwickelten Konzeptualisierungen. Andererseits nimmt das Projekt die eingangs skizzierte Fortentwicklung der Ansätze zu Lehrerkompetenzen auf, indem handlungsbezogene Facetten von Lehrerkompetenzen eine zentrale Rolle spielen, hier die Situationswahrnehmung und die Handlungsplanung und deren Zusammenhang mit mathematikdidaktischem Wissen. Die Nichtberücksichtigung von erziehungswissenschaftlichem Wissen ist angesichts der hohen pädagogischen Anforderungen in diesem Bereich eine Lücke, die in Folgestudien zu schließen sein wird. Die Wissensfacetten werden traditionell im Papier-und-Bleistift-Format erhoben, die Situationswahrnehmung und die Handlungsplanung videobasiert. Sowohl die Konzeptualisierung als auch die Anlage der Studie greifen die anfangs erwähnten Ansätze zu Lehrerkompetenzen auf und adaptieren sie geeignet auf eine neue Berufsgruppe. Damit erscheint die Studie geeignet, insgesamt die Diskussion zum Professionswissen in pädagogischen Berufen voranzutreiben.

Bouley et al. nehmen mit ihrem Beitrag „Der Einfluss von universitären und außeruniversitären Lerngelegenheiten auf das Fachwissen und fachdidaktische Wissen von angehenden Lehrkräften an kaufmännisch-berufsbildenden Schulen“ eine weitere bisher weniger erforschte Gruppe von Lehrkräften in den Fokus, nämlich angehende Lehrkräfte an kaufmännisch-berufsbildenden Schulen. Dabei knüpfen sie – wie die bereits diskutierten Projekte – an die aktuelle Diskussion zu Lehrerkompetenzen und dem Ansatz von Shulman an. Wie das Projekt von Riese et al. zu angehenden Physiklehrkräften wird nicht die Shulman’sche Trias untersucht, sondern die Autor(inn)en konzentrieren sich auf Fachwissen und fachdidaktisches Wissen, was angesichts einer leistungsmäßig sehr heterogenen Schülerschaft mit sehr unterschiedlichem soziokulturellem Hintergrund in Folgestudien erweitert werden sollte. Die Studie untersucht die Zusammenhänge des Professionswissens mit den Lerngelegenheiten, was eine neue Facette in die Diskussion einbringt, nämlich die Auswirkungen von Lerngelegenheiten auf Lehrerkompetenzen; insbesondere wird der Einfluss des außeruniversitär gewonnenen Vor-

wissens – z. B. in Form einer vor dem Studium absolvierten Ausbildung an einer kaufmännischen Vollzeitschule oder einer kaufmännischen Berufsausbildung – untersucht. Wie in dem Projekt von Riese et al. erfolgt auch hier eine thematische Einschränkung auf einen zentralen Bereich der kaufmännisch-berufsbildenden Schulen, nämlich das Rechnungswesen.

Der im Projekt deutlich gewordene hohe Einfluss außeruniversitärer Lerngelegenheiten auf das Fachwissen und das fachdidaktische Wissen bestätigt einerseits die hohe Bedeutung von Praxiserfahrungen – als solche ist ja eine kaufmännische Berufsausbildung anzusehen –, worauf immer wieder in der einschlägigen Diskussion hingewiesen wird. Zum Weiteren wird aber auch deutlich, wie heterogen die Ausbildung an den einzelnen Universitäten ist, sodass ein von den Anforderungen der Berufspraxis konstruiertes Testinstrument deutlich an seine Grenzen stößt. Insgesamt wird deutlich, dass die Erhebung von Lehrerprofessionswissen in diesen sogenannten gering strukturierten Domänen (Blömeke et al., 2011), zu denen auch die Ausbildung von Berufsschullehrkräften an kaufmännisch-berufsbildenden Schulen gehört, sowohl vom disziplinären Verständnis dieser Wissenschaften als auch von ihrer Umsetzung in der universitären Praxis an ihre Grenzen stößt.

Der Beitrag von Zlatkin-Troitschanskaia et al. zum „Erwerb wirtschaftswissenschaftlicher Fachkompetenz im Studium – Eine mehrebenenanalytische Betrachtung von hochschulischen und individuellen Einflussfaktoren“ erweitert den Fokus der bisherigen Beiträge weg von der Lehrerbildung hin zur Fachkompetenz im Studium. Es wird untersucht, inwieweit das Konstrukt „wirtschaftswissenschaftliche Kompetenz von Studierenden“ durch hochschulische und individuelle Einflussfaktoren beeinflusst wird. Die Studie knüpft ebenfalls an die kompetenzbezogene Forschung, wie sie für die Lehrerbildung eingangs beschrieben wurde, an. Obwohl es Konsens ist, dass kognitive Dispositionen auch Überzeugungen und Einstellungen beinhalten, konzentriert sich die einschlägige Forschung in diesem Bereich auf Fachwissen. In Einklang mit der Studie von Bouley et al. wird deutlich, dass Studierende, die z. B. eine kaufmännisch-verwaltende Berufsausbildung absolviert haben, Vorteile haben gegenüber Studierenden ohne eine solche Ausbildung. Damit wird nicht nur die hohe Bedeutung von Praxiserfahrungen deutlich, sondern es wird auch gezeigt, dass es im Hochschulbereich nicht oder kaum gelingt, die Unterschiede in den Eingangsvoraussetzungen unter einer Geschlechter- bzw. Mehrsprachigkeitsperspektive abzubauen.

Insgesamt machen die Beiträge deutlich, welche Breite die aktuelle Diskussion zu Kompetenzen von Studierenden im universitären Bereich abdeckt. Es wird aber auch deutlich, dass trotz vielfältiger fachspezifischer Unterschiede aktuelle Ansätze zu berufsbezogenen Kompetenzen von Studierenden einen großen gemeinsamen Kern aufweisen, wobei die hier versammelten Beiträge an unterschiedliche Phasen und Aspekte der aktuellen Debatte zur Lehrerbildung anknüpfen. Die Implementierung der in der aktuellen internationalen Diskussion zu Lehrerkompetenzen als Konsens anzusehenden Berücksichtigung der Shulman’schen Trias der Domänen Fachwissen, fachdidaktisches Wissen und erziehungswissenschaftliches Wissen bleibt eine Herausforderung für die Zukunft.

Literatur

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

Blömeke, S., Bremerich-Vos, A., Haudeck, H., Kaiser, G., Lehmann, R., Nold, G., Schwippert, K., & Willenberg, H. (Hrsg.) (2011). *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen: Erste Ergebnisse aus TEDS-LT*. Münster: Waxmann.

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

Blömeke, S., & Delaney, S. (2012). Assessment of teacher knowledge across countries: A review of the state of research. *ZDM – The International Journal of Mathematics Education, 44*, 223–247.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (im Druck). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*.

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

Blömeke, S., Kaiser, G., & Lehmann, R. (Hrsg.) (2010b). *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte der Sekundarstufen im internationalen Vergleich.* Münster: Waxmann.

Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M., & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf – Fachbezogenes Wissen als Voraussetzung für Wahrnehmung, Interpretation und Handeln im Unterricht. *Zeitschrift für Erziehungswissenschaft, 14*(3), 509–542 [DOI: 10.1007/s11618-014-0564-8].

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

Bromme, R. (1992). *Der Lehrer als Experte*. Bern: Huber.

Buchholtz, N., & Kaiser, G. (2013). Improving Mathematics Teacher Education in Germany: Empirical Results from a Longitudinal Evaluation of Innovative Programs. *International Journal of Science and Mathematics Education, 11*(3), 949–977.

Buchholtz, N., Kaiser, G., & Blömeke, S. (2014). Die Erhebung mathematikdidaktischen Wissens – Konzeptualisierung einer komplexen Domäne. *Journal für Mathematikdidaktik, 35*(1), 101–128.

König, J., & Seifert, A. (Hrsg.) (2012). *Lehramtsstudierende erwerben pädagogisches Professionswissen: Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung.* Münster: Waxmann.

Kunina-Habenicht, O., Schulze-Stocker, F., Kunter, M., Baumert, J., Leutner, D., Förster, D., Lohse-Bossenz, H., & Terhart, E. (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., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*, 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Research, 57*, 1–22.

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

Anschrift der Autorin

Prof. Dr. Gabriele Kaiser, Universität Hamburg,
Arbeitsbereich Mathematikdidaktik,
Von-Melle-Park 8, 20146 Hamburg, Deutschland
E-Mail: gabriele.kaiser@uni-hamburg.de

Kati Trempler/Andreas Hetmanek mit Christof Wecker/Jan Kiesewetter/
Mia Wermelt/Frank Fischer/Martin Fischer/Cornelia Gräsel

Nutzung von Evidenz im Bildungsbereich

Validierung eines Instruments zur Erfassung von Kompetenzen der Informationsauswahl und Bewertung von Studien

Zusammenfassung: Evidenzbasierte Praxis kann nach dem Vorbild der Medizin als reflektierte Nutzung der besten verfügbaren empirischen Befunde für die berufliche Tätigkeit beschrieben werden. Dies stellt eine komplexe Anforderung für pädagogisches Personal dar. Ziel der vorliegenden Arbeit ist die Validierung eines Instruments zur Erfassung der dafür benötigten Kompetenz. Im Fokus stehen die beiden Teilkompetenzen *Informationsauswahl* und *Bewertung von Studien.* An der Studie nahmen 341 Studierende aus erziehungswissenschaftlichen Studiengängen teil. Es wurde ein fallbasierter Online-Test für Informationsauswahl und Bewertung von Studien durchgeführt. Die Ergebnisse liefern Hinweise für die Trennbarkeit der beiden Teilkompetenzen und zeigen plausible Zusammenhänge zu weiteren Variablen.

Schlagworte: Validierung, Kompetenzmessung, fallbasierte Entscheidung, evidenzbasierte Praxis im Bildungsbereich, Primärstudien

1. Einleitung

Die alltäglichen Anforderungssituationen, mit denen Praktiker im Bildungsbereich konfrontiert werden, sind vielfältig und umfassen die Organisation und Aufrechterhaltung von Aktivitäten innerhalb sozialer Gruppen, die Entwicklung und Verbreitung von Inhalten sowie die zeitliche Planung von Abläufen (Bromme, 2014, S. 89). Zahlreiche leicht zugängliche ratgebende und anleitende Materialien zur Gestaltung dieser Situationen existieren; vor allem in Form von Zeitschriften, Handbüchern und im Internet. Qualitativ hochwertige und wissenschaftlich abgesicherte Literatur allerdings ist selten darunter zu finden. Dabei kann sie durchaus entscheidend zur Bewältigung von Anforderungen der beruflichen Praxis beitragen (Bromme, Prenzel, & Jäger, 2014, S. 8). Relevante wissenschaftliche Forschungsarbeiten zu finden, auszuwählen sowie mit Blick auf eine konkret anstehende Entscheidung zu bewerten und nutzbar zu machen, stellt eine komplexe Anforderung für pädagogisches Personal dar. Eine reflektierte und auf-

geklärte evidenzbasierte Praxis im Bildungsbereich kann sich aber nur dann entwickeln, wenn diese Anforderung von verantwortlichen Akteuren bewältigt wird. Die *Informationsauswahl* und die *Bewertung von Studien* sind zwei wichtige Bestandteile der komplexen Kompetenz des evidenzbasierten Argumentierens. Im Mittelpunkt dieses Beitrages stehen die theoretische Modellierung und die Validierung eines Tests für die beiden Teilkompetenzen.

2. Hintergrund

2.1 Evidenzbasierung im Bildungsbereich

Ihren Ursprung hat die Evidenzbasierung im Bereich der Medizin. Als neues Paradigma der medizinischen Ausbildung und Patientenversorgung wurde bereits vor mehr als zwei Jahrzenten durch die Evidence-Based Medicine Working Group (1992) die evidenzbasierte Praxis eingeführt. Die Idee, versorgungsrelevante Fragen stärker auf Grundlage vorhandener wissenschaftlicher Erkenntnisse in Form von wissenschaftlicher Primärliteratur und weniger auf Grundlage individueller bzw. tradierter Erfahrung zu bearbeiten, wird zur Erfolgsgeschichte (u. a. Montori & Guyatt, 2008, S. 1814). Erste Versuche, Unterricht evidenzbasiert zu gestalten, finden sich konsequenterweise in der medizinischen Ausbildungsforschung (Harden, Grant, Buckley & Hart, 1999).

Seit der Veröffentlichung der ersten PISA-Ergebnisse 2000 (Baumert et al., 2002) wird ein ähnlicher Paradigmenwechsel auch für den Bildungsbereich diskutiert (u. a. Bromme et al., 2014; Slavin, 2008) und teilweise aktiv politisch vorangetrieben (z. B. KMK, 2014, S. 3). Im Mittelpunkt der Diskussion steht die Frage, in welcher Weise bildungswissenschaftliche Befunde für die Praxis gesucht, beurteilt und verwendet werden können – sowohl auf der Ebene politischer Steuerung wie auch auf der Ebene von Entscheidungen einzelner Personen (Bromme et al., 2014, S. 6). Damit die Forderung nach evidenzbasierter Praxis im Bildungsbereich nach dem Vorbild der Medizin fruchtbar werden kann, muss ein kompetenter Umgang mit originär wissenschaftlicher Literatur als Grundvoraussetzung für eine evidenzbasierte Entscheidungskultur durch die im Bildungsbereich tätigen Praktiker (z. B. Erzieherinnen und Erzieher, Lehrkräfte an allen Schulformen, Fort- und Weiterbildner) gegeben sein. Hierfür muss der Umgang mit wissenschaftlicher Literatur in das professionelle Berufsverständnis integriert werden (Tugwell, Haynes & Sackett, 1992).

Für den Bildungsbereich wird evidenzbasierte Praxis nach dem Vorbild der Medizin als reflektierte Nutzung der besten verfügbaren und empirisch gesicherten Evidenz für das Treffen und die Begründung von beruflichen Entscheidungen definiert (vgl. Bromme et al., 2014, S. 9; Sackett, Rosenberg, Gray, Haynes & Richardson, 1996, S. 71). Damit stellt sie einen Gegenentwurf zur verbreiteten Praxis dar, Entscheidungen überwiegend auf Basis von individuellen Erfahrungen, Ideologien oder subjektiven Überzeugungen zu treffen (Harden et al., 1999; Slavin, 2008). Diskussionen zur Evidenzbasierung im Bildungsbereich befassen sich bislang vor allem mit der Ebene der evidenzbasierten

Steuerung, also der Nutzung wissenschaftlicher Evidenz als Grundlage für bildungspolitische Entscheidungen (*evidence-based reform;* Slavin, 2008). Eine weitere Ebene der Nutzung von Evidenz als Grundlage für individuelles professionelles Handeln im Bildungsbereich findet bislang wenig Beachtung (Bromme et al., 2014), obwohl beispielsweise bereits im Jahr 2004 durch die Kultusministerkonferenz die Standards für die Lehrerbildung vorgelegt wurden, in denen die Planung, Organisation und Reflexion von Lehr-Lern-Prozessen anhand bildungswissenschaftlicher Erkenntnisse als „Kernaufgabe" von Lehrkräften definiert wird (KMK, 2014, S. 3). Diese Kernaufgabe erfordert Fähigkeiten der Rezeption und Bewertung bildungswissenschaftlicher Forschungsergebnisse sowie der Nutzbarmachung von Ergebnissen der Bildungsforschung für die Tätigkeit als Lehrkraft (KMK, 2014, S. 13). Bildungswissenschaftliche Evidenz umfasst empirisch abgesicherte Erkenntnisse, die aus qualitativ hochwertigen Untersuchungen der Bildungsforschung gewonnen werden; sie ermöglicht es, Hinweise für die Wirksamkeit pädagogischer Maßnahmen zu extrahieren und pädagogische Entscheidungen auf der Grundlage von abgesichertem Wissen zu rechtfertigen (Cook, Smith & Tankersley, 2012, S. 498 ff.). Welche Art von Forschung hierfür besonders geeignet ist, bildet Gegenstand fortlaufender Diskussion (Beelmann, 2014; Cook & Gorard, 2007; Fischer, Waibel & Wecker, 2005; Wecker, 2013). Auch wenn systematische Forschungssynthesen in der Regel höheren Qualitätsansprüchen genügen (siehe Diskussion bei Beelmann, 2014), sind empirische Forschungsartikel die primäre Quelle für forschungsbasierte Erkenntnisse: Zum einen gibt es zu vielen relevanten Themenkomplexen im Bildungsbereich bislang kaum systematische Reviews oder Metaanalysen, zum anderen gehen in derartigen Überblicksarbeiten notwendigerweise viele umsetzungsrelevante Details verloren. Damit ist die Fähigkeit, kritisch und kompetent mit empirischer Primärliteratur umzugehen, als von grundsätzlicher Bedeutung für evidenzbasierte Praxis anzusehen (EBMWG, 1992). Aus diesem Grund befasst sich die vorliegende Untersuchung mit dem Umgang von Personen im Bildungsbereich mit empirischen Studien.

2.2 Die Kompetenz im evidenzbasierten Argumentieren

Im Folgenden werden zunächst anhand einer alltäglichen Anforderungssituation im Bildungsbereich zwei mögliche Handlungsalternativen vorgestellt, die den Paradigmenwechsel von aktuell gängigen Verfahren der professionellen pädagogischen Praxis zu einer stärker evidenzbasierten Praxis im Bildungsbereich skizzieren. Daran anschließend wird das im Rahmen des hier vorgestellten Forschungsvorhabens entwickelte Prozessmodell für Leistungen beim evidenzbasierten Argumentieren vorgestellt. Genauere Spezifikationen des Zusammenspiels der beiden untersuchten Teilkompetenzen mit weiteren Komponenten bzw. Faktoren und das Vorgehen bei der Validierung des Testinstrumentes zur Erfassung der Leistungen werden im Anschluss daran präsentiert.

Die Anforderungssituation

Alltagsszenario: zu einer der häufigsten Aufgaben von pädagogischen Praktikern im Bildungsbereich gehört die Planung und Anleitung von Aktivitäten zum Erwerb von Wissen und Fertigkeiten. Die Frage ist nun, wie der Einzelne dieser Anforderung begegnet. Wie geht er bei der Entwicklung eines effektiven Konzepts für Wissensvermittlung bzw. Fertigkeitserwerb vor?

Der aktuell gängige Weg: Der pädagogische Praktiker denkt über den Inhalt bzw. das Lehrziel nach und überlegt, wie er es am besten erreichen könnte. Zusätzlich sichtet er alte Konzepte mit ähnlichen Inhalten bzw. Zielen, spricht eventuell mit weiteren Kolleginnen, konsultiert einen „methodischen Werkzeugkasten" im Internet oder in Buchform und blättert – sofern genug Zeit zur Verfügung steht – in den Unterlagen aus dem eigenen Studium oder Materialien aus der Ausbildung. Dabei entsteht ein Vermittlungskonzept, das überwiegend inhaltsfokussiert ist und methodisch auf unsystematischen Beobachtungen und Erfahrungen sowie Intuition aufbaut.

Der Weg der evidenzbasierten Praxis: Der verantwortliche Pädagoge analysiert den zu vermittelnden Lehrinhalt, die Zielgruppe sowie weitere wichtige Einflussgrößen. Insofern ihm noch kein evidenzbasiertes Wissen zu diesem Typ von Anforderungssituation zur Verfügung steht, überlegt er sich Suchbegriffe, unter denen wissenschaftliche Erkenntnisse zu möglichst ähnlichen Typen von Lerninhalten bei ähnlicher Ausgangslage publiziert sein könnten, und sucht systematisch in Fachportalen nach passenden empirischen Untersuchungen, wählt die wichtigsten (relevantesten und qualitativ besten) aus und liest und bewertet sie systematisch anhand eines elaborierten Bewertungsschemas. Dabei achtet er bspw. auf die Passung der Studie zu seinen Lehrzielen und den gegebenen Bedingungen sowie auf die Qualität der Studien. Auf Grundlage dieser Informationen entwickelt er sein Vermittlungskonzept. Bei der Dokumentation zieht er die gewählte Primärliteratur als Begründung für das gewählte Vorgehen heran. Bei den zu treffenden Entscheidungen kommt systematisch abgesichertem Wissen überwiegendes Gewicht zu.

So könnte sich der Paradigmenwechsel im Bildungsbereich nach dem Vorbild der Medizin vollziehen. In der hier bewusst überzeichneten „Reinform" ist und bleibt diese Art der Evidenzbasierung Fiktion – in der Medizin gleichermaßen wie im Bildungsbereich. In beiden Bereichen werden aktuell lebhafte Debatten um eine sinnvolle, praktizierbare und zielführende Ausgestaltung geführt (Pant, 2014, S. 81).

In jedem Fall wird für die Umsetzung einer stärker evidenzbasierten Praxis vom professionellen Praktiker ein ganzes Bündel von neuen Fertigkeiten verlangt. Besonders grundlegend sind die effiziente Literatursuche und -auswahl sowie die Anwendung von Regeln zur Bewertung von empirischer Forschung (EBMWG, 1992, S. 2420). Weiterhin müssen die abgesicherten Erkenntnisse aus der Forschung mit anderen Faktoren (Praktikabilität, Erfahrungen mit der spezifischen Gruppe von Lernenden etc.) integriert und bei der Entscheidung berücksichtigt sowie reflektiert, dokumentiert und oftmals auch kommuniziert werden. Dieser Ablauf stellt eine komplexe und vielschichtige kognitive Anforderung dar, für deren Bewältigung verschiedene Teilfertigkeiten, spezifisches Wissen und Einstellungen zusammenspielen und verschiedene Phasen kognitiver

Aktivität durchlaufen werden müssen. Als Bezeichnung für diesen Problemlöseprozess führen wir den Begriff „evidenzbasiertes Argumentieren" ein. Beim evidenzbasierten Argumentieren und den spezifischen Anforderungen der *Informationsauswahl* und *Bewertung von Studien* handelt es sich um gering strukturierte Probleme (Simon, 1973). Die Entwicklung eines theoretischen Modells für den beschriebenen Prozess ist eine zentrale Leistung des zugrunde liegenden Forschungsvorhabens.

Zur theoretischen Modellierung der Kompetenz im evidenzbasierten Argumentieren im Bildungsbereich greifen wir Konzeptionen aus der Medizin (Rosenberg & Donald, 1995, S. 1123) auf und erweitern diese mit Theorien und Erkenntnissen aus der Forschung zum *Information Problem Solving* (im Folgenden „IPS"; Brand-Gruwel, Wopereis & Vermetten, 2005) für den Bereich der systematischen Informationsbeschaffung, -bewertung und -integration und zum *Evidence-based Reasoning Framework* (u. a. Brown, Furtak, Timms, Nagashima & Wilson, 2010) für den Teilbereich der Entscheidungsfindung und argumentativen Begründung.

Aufgrund der besonders grundlegenden Bedeutung konzentrieren wir uns in diesem Beitrag auf die beiden Teilbereiche (1) *Informationsauswahl* und (2) *Bewertung von Studien.*

Informationsauswahl

Der erste Schritt zur Bewältigung der Anforderungen der evidenzbasierten Praxis ist das Suchen und Auswählen aussagekräftiger Untersuchungen, um ein Informationsproblem zu lösen. Dieser Schritt wird üblicherweise anhand von Datenbanken oder in Internetsuchmaschinen durchgeführt. Von Brand-Gruwel, Wopereis und Walraven (2009) wurde ein ausführliches Modell zur Lösung von Informationsproblemen vorgelegt und empirisch bestätigt. Zentraler Bestandteil dieses Modells sind die iterativen Prozesse von Informationssuche, Überfliegen und tiefer Verarbeitung der Information (S. 1208). Für das Gelingen dieser Prozesse ordnen die Autoren den regulierenden Aktivitäten Orientierung, Monitoring, Steuerung und Prozessevaluation eine zentrale Rolle zu (S. 1209). Experten-Novizen-Vergleiche zeigen, dass Experten deutlich mehr Zeit mit vertiefter Beschäftigung mit dem Inhalt verbringen als Novizen (Brand-Gruwel et al., 2005, S. 498). Dies deckt sich mit Erkenntnissen aus einem weiteren relevanten Forschungszweig zum Verständnis multipler Quellen (*Multiple Source Comprehension;* u. a. Wiley et al., 2009). Auch hier spielen das Verstehen und die Bewertung der Relevanz und Nützlichkeit einer Information eine zentrale Rolle (Goldman, Lawless, Pellegrino, Braasch & Gomez, 2012, S. 182). Diese Reflexion und Bewertung sind insbesondere notwendig bei einer gegebenen Fülle an Informationen, die beispielsweise bei einer Internetsuche in Form von Listen ungefiltert – d. h. ohne jegliche Form der Plausibilitäts- und Qualitätsprüfung – ausgegeben werden. Für derartige Tätigkeiten sind spezifische Wissensbestände zur Bewertung der Relevanz und Qualität der Informationen notwendig, damit erfolgreich die aussagekräftigsten Forschungserkenntnisse aufgefunden werden können. Die Wichtigkeit von anforderungsspezifischem Vorwissen sowie epistemologischen Überzeugungen wird auch an anderer Stelle betont (Lazonder & Rouet, 2008) und findet in einem systematischen Re-

view der Forschungsarbeiten zum IPS Unterstützung (Wopereis & van Merriënboer, 2011, S. 234–235).

Ausgehend von diesen Vorarbeiten modellieren wir die Informationsauswahl im evidenzbasierten Argumentieren als Prozess von Suche, Überfliegen und Berücksichtigung weiterer Informationen, bei dem darüber hinaus anforderungsspezifisches Vorwissen sowie epistemologische Überzeugungen eine bedeutsame Rolle spielen. Als anforderungsspezifische Wissensbestände nehmen wir Grundkenntnisse (Terminologie und Zusammenhänge) aus der pädagogischen und psychologischen Lehr-Lern-Forschung sowie zur empirischen Forschungsmethodik an.

Bewertung von Studien

Nachdem wissenschaftliche Forschungsarbeiten auf Grundlage von schnell überblickbaren Informationen (bspw. Schlagwörter und Zusammenfassungen in einer Literaturdatenbank) als relevant eingestuft und ausgewählt wurden, müssen die Publikationen einer systematischen Beurteilung unterzogen werden. Für den medizinischen Bereich haben Harden et al. (1999) ein Bewertungsraster entwickelt, mit dem sechs Aspekte wissenschaftlicher Untersuchungen abgedeckt werden: Qualität, Übertrag- und Anwendbarkeit des Behandlungsansatzes auf den aktuellen Fall, Reichweite, Stärke und Validität der Befunde sowie Vergleichbarkeit der Rahmenbedingungen. Für den Bildungsbereich fehlt bislang ein derartiges Beurteilungsschema.

Zur Entwicklung eines für den Bildungsbereich adäquaten Bewertungsschemas haben wir das Schema aus der Medizin als Ausgangspunkt herangezogen und durch Qualitätsaspekte aus der pädagogisch-psychologischen Methodenlehre erweitert und geschärft. Zu diesem Zweck haben wir zusätzlich die vier Aspekte von Validität nach Cook und Campbell (1979) sowie das UTOS-Bewertungsschema zur Übertragbarkeit von Studienergebnissen nach Cronbach (1982) integriert. Dabei entstand ein Beurteilungsschema zur systematischen Bewertung von Relevanz für eine konkrete Anwendungssituation und allgemeiner methodischer Qualität einer wissenschaftlichen Studie (siehe Abb. 3).

Der pädagogische Praktiker wird bei der (systematischen) Bewertung von Studien mit einer komplexen Anforderung konfrontiert. Zur Bewältigung muss ein komplizierter kognitiver Prozess realisiert werden, der das Lesen und Verstehen der Studie sowie letztlich das Beurteilen der einzelnen Dimensionen der wissenschaftlichen Arbeit umfasst. Es ist plausibel anzunehmen, dass umfangreiche inhaltliche und forschungsmethodische Wissensbestände im Zusammenspiel mit epistemologischen Überzeugungen und der Motivation zentral für die Bewältigung der Aufgabe sind.

Hintergrundvariablen

Grundsätzlich werden kognitive Leistungen von *allgemeinen kognitiven Fähigkeiten* (u. a. Amelang, 2000) und *spezifischem Wissen* (Feltovich, Prietula & Ericsson, 2006) beeinflusst, und Leistungsunterschiede bei komplexen kognitiven Aufgaben wie dem evidenzbasierten Argumentieren können teilweise durch Unterschiede in diesen beiden Variablen erklärt werden (Klieme, Funke, Leutner, Reimann & Wirth, 2001, S. 189;

Mayer, 2003, S. 263 ff.). Die Anforderungen bei der *Informationsauswahl* und der *Bewertung von Studien* können zudem nur bewältigt werden, wenn anforderungsspezifisches – also pädagogisch-psychologisches und forschungsmethodisches – Wissen in ausreichendem Maß vorhanden ist sowie eine ausreichende Ausprägung insbesondere verbaler Intelligenz gegeben ist.

Die Struktur der Anforderungen an den evidenzbasiert agierenden Praktiker legt es nahe, dass einige kognitive Aktivitäten Überlappungen mit denjenigen beim *wissenschaftlichen Denken* haben. Wissenschaftliches Denken umfasst die bewusste und kontrollierte Anwendung wissenschaftlicher Methoden und Prinzipien für die Koordination von Theorie und wissenschaftlichen Befunden in Argumentations- und Problemlösesituationen (Zimmerman, 2000, S. 114). Leistungsunterschiede in den Bereichen *Informationsauswahl* und *Bewertung von Studien* könnten daher möglicherweise ebenfalls durch Unterschiede in der Ausprägung dieser Fertigkeit erklärt werden.

Dies gilt auch für die kognitiven Prozesse bei der Prüfung der Zuverlässigkeit und Genauigkeit einer Aussage, der Unterscheidung zwischen relevanten und nicht relevanten Informationen sowie beim wissenschaftlich-analytischen Schlussfolgern. Daher nehmen wir an, dass Fertigkeiten des *kritischen Denkens* (Astleitner, Brünken & Zander, 2002) eine weitere relevante Einflussgröße beim Bewältigen von Anforderungen an den evidenzbasierten Praktiker darstellen.

Es wurde bereits herausgearbeitet, dass auch persönliche Überzeugungen eine Rolle spielen; so ist anzunehmen, dass *epistemologische Überzeugungen* zu Leistungen im evidenzbasierten Argumentieren beitragen. Sie werden beschrieben als „(…) intuitive Theorien, die die Art der Begegnung mit der erkennbaren Welt vorstrukturieren“ (Baumert & Kunter, 2006, S. 498). Diese Einstellungen zu Wissen und Wissenserwerb beeinflussen das Lernen, Denken und Schlussfolgern (Bromme et al., 2014; Hofer, 2000; Hofer & Pintrich, 1997). Auch die subjektive Einschätzung der Vertrautheit mit forschungsmethodischen Begriffen, also *Überzeugungen zum eigenen forschungsmethodischen Wissen,* sollte Unterschiede in Leistungen bei *Informationsauswahl* und *Bewertung von Studien* im Kontext von evidenzbasierter Praxis anteilig erklären können. Diese Überzeugungen umfassen Einschätzungen der eigenen Fähigkeiten des Umgangs mit forschungsmethodischen Inhalten und könnten sich ebenfalls in Leistungen des evidenzbasierten Argumentierens niederschlagen.

Da bislang kaum Erkenntnisse zu Einflussgrößen im Kontext gering strukturierter Probleme (Simon, 1973) beim Erbringen von Leistungen in diesem Zusammenhang vorliegen, sind Vorhersagen exakter (kausaler) Zusammenhänge und Abhängigkeiten nicht möglich.

2.3 Vorgehensweise bei der Validierung

Inhaltsvalidierung: Belege für die Inhaltsvalidität werden in aller Regel nicht quantitativ empirisch, sondern theoretisch erbracht und beziehen sich darauf, ob die Inhalte von Items das untersuchte Attribut abbilden (vgl. Hartig, Frey & Jude, 2012). Die Inhalts-

validierung umfasst in der Regel die Begutachtung der konzipierten Items eines Tests durch Experten, deren Expertise systematisch nachgewiesen werden sollte (Jenßen, Dunekacke & Blömeke, 2015, in diesem Beiheft). Wir betrachten die Inhaltsvalidität dann als gegeben, wenn die befragten Experten in ihrer Bearbeitung der Aufgaben zur Informationsauswahl sowie der Bewertung von Studien eine hohe Übereinstimmung aufweisen. Bei den Aufgaben zur *Informationsauswahl* ist die Übereinstimmung dann gegeben, wenn Experten sich über die Relevanz von Studien einig sind, d. h. dieselben Studien als relevant, weniger relevant und nicht relevant einschätzen. Bei der *Bewertung von Studien* kann von Übereinstimmung gesprochen werden, wenn Experten Studien anhand der oben genannten Qualitätsaspekte kongruent bewerten.

Konstruktvalidierung: Im Sinne von Cronbach und Meehl (1955) dient die Konstruktvalidierung dem Nachweis, welche latente Variable (Kompetenz) das Kriterium (Performanz) bestimmt. Prozeduren für die Bestimmung der Konstruktvalidität (nach Cronbach & Meehl, 1955, S. 287–288) umfassen unter anderem (1) Korrelationen zwischen den einzelnen Bereichen innerhalb der Gesamtkompetenz (Korrelation zwischen den Bereichen *Informationsauswahl* und *Bewertung von Studien*) und (2) Aussagen über die interne Struktur bzw. Konsistenz der Teilkompetenzen, welche durch Inter-Item-Korrelationen nachgewiesen werden kann (interne Konsistenz der Bereiche *Informationsauswahl* und *Bewertung von Studien*). Aufgrund der inhaltlichen Nähe (überlappende kognitive Aktivitäten) und klaren theoretischen Unterschiedenheit werden Maße der Intelligenz, spezifisches Wissen, wissenschaftliches und kritisches Denken, epistemologische Überzeugungen und Überzeugungen zum forschungsmethodischen Wissen zur Validierung herangezogen. Durch die Analyse von Zusammenhängen der benannten Variablen mit Leistungsmaßen für die Bereiche *Informationsauswahl* und *Bewertung von Studien* können weitere Argumente für die Konstruktvalidität erbracht werden.

2.4 Zielsetzung und Fragestellung der vorliegenden Studie

Ziel der vorliegenden Studie war es, Testinstrumente zur Messung der Leistung bei *Informationsauswahl* und *Bewertung von Studien* im Rahmen des evidenzbasierten Argumentierens zu entwickeln und zu validieren. Nach der Darstellung des Instrumentes und insbesondere der Operationalisierung der Teilbereiche *Informationsauswahl* und *Bewertung von Studien* werden Argumente für die Validität zusammengetragen (Kane, 1992). Die Fragestellung des vorliegenden Beitrags lautet:

Welche empirischen Argumente können für die Inhalts- und Konstruktvalidität des entwickelten Verfahrens zur Messung der *Informationsauswahl* und der *Bewertung von Studien* im Kontext von evidenzbasiertem Argumentieren angeführt werden?

3. Methode

3.1 Erhebungsinstrumente

Szenario des Kompetenztests
Für die Erfassung der Kompetenz im evidenzbasierten Argumentieren wurde ein computerbasiertes Szenario entwickelt. Computerbasierte Szenarien bieten im Gegensatz zu Papier-und-Bleistift-basierten Instrumenten den Vorteil, komplexe Interaktionsmuster messbar und quantifizierbar zu machen sowie die Messung von komplexen Kompetenzen in kontextualisierten und möglichst realitätsnahen Problemlösesituationen durchzuführen (siehe Diskussion bei Frey & Hartig, 2013, S. 54). Innerhalb des Szenarios sollen die Probanden (in der Rolle einer Lehrkraft) im Rahmen eines wirtschaftsfinanzierten „business start"-Wettbewerbs, in dem Schüler ein eigenes Geschäftsmodell entwickeln und präsentieren lernen sollen, ein Präsentationstraining für eine Gruppe mit 15-jährigen Schülerinnen und Schülern vorbereiten. Im Laufe der Trainingsplanung gilt es, zwei Entscheidungen zu treffen: (1) Die Einleitung und Eröffnung einer Präsentation soll mit Videobeispielen trainiert werden. Die Probanden sollen entscheiden, ob allein vorbildliche oder eine Kombination von vorbildlichen und fehlerhaften Videobeispielen eingesetzt werden sollten. (2) Im zweiten Teil des Trainings sollen Grundlagen aus der Rhetorik vermittelt werden. Hier ist von den Probanden zu entscheiden, ob in diesem Fall die Methode des „Gruppenpuzzles" oder die Einzelarbeit eingesetzt werden sollte. Der fiktive Geldgeber für das Trainingsprogramm besteht auf Evidenzbasierung; daher ist es notwendig, nach geeigneten wissenschaftlichen Studien zu suchen und insbesondere deren Qualität zu bewerten.

Zu beiden Entscheidungen liegen ausreichend Forschungsergebnisse vor, sodass bei der Entwicklung der Materialien aus einer Menge von wissenschaftlichen Studien ausgewählt werden konnte. Zudem sind die Inhaltsbereiche für unterschiedliche pädagogische Settings relevant (z. B. für angehende Lehrpersonen an Schulen oder auch für angehende Weiterbildnerinnen und Weiterbildner). Die ausgewählten Studien wurden in Bezug auf Fragestellungen, Methodik und Ergebnisse nicht verändert. Es wurden strukturierte Kurzfassungen der Originalarbeiten erstellt, die sich in einer anderen Studie als leichter zugänglich und besser verständlich erwiesen haben (Hetmanek et al., 2014). Alle Materialien wurden in deutscher Sprache präsentiert.

Die aus den theoretischen Vorüberlegungen abgeleiteten Teilschritte für die evidenzbasierte Entscheidungsfindung, die in unserer Testumgebung entsprechend umgesetzt wurden, sind in Tabelle 1 (inkl. Bearbeitungszeiten) dargestellt.

Pilotierungen ergaben, dass nach der Bearbeitung von zwei Texten die Sicherheit und Vertrautheit mit dem Bewertungsraster deutlich zunimmt. Daher wurde den Probanden vor der Bearbeitung der eigentlichen Testaufgabe eine Übungsentscheidung vorgelegt, anhand derer die Bewertung von Studien an zwei wissenschaftlichen Texten erprobt werden konnte.

Im Rahmen dieses Beitrags fokussieren wir die Instrumente für die Erfassung der Teilkompetenzen 1b, 1c und 2 (siehe Tab. 1).

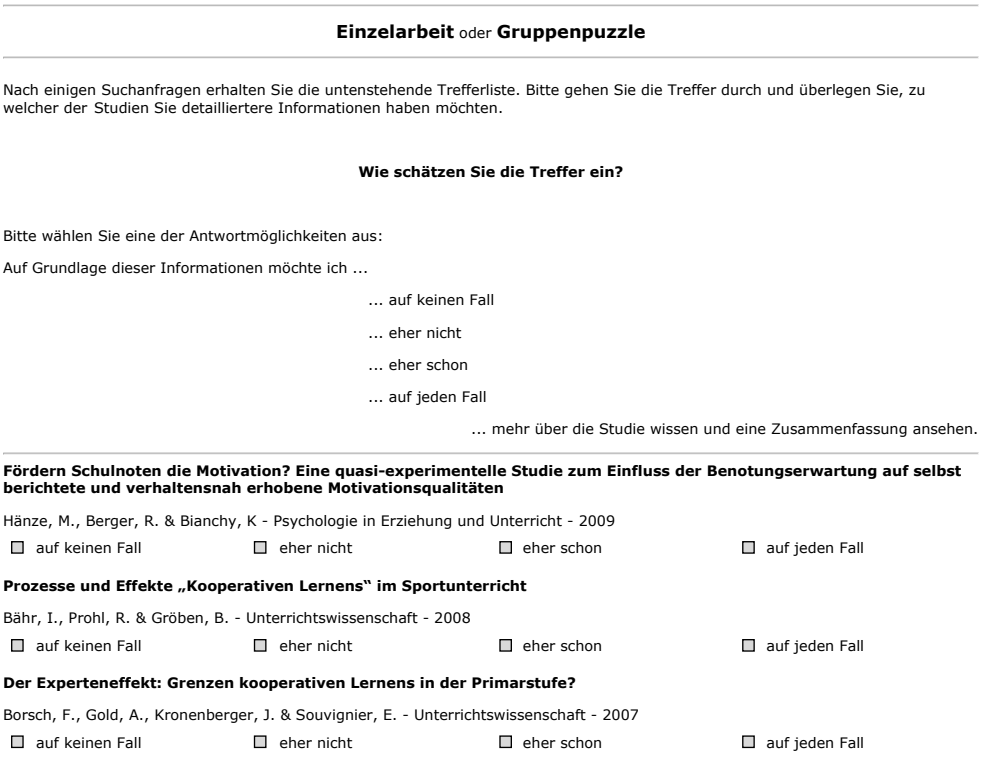
	Teilkompetenz	vorgesehene Zeit
1a	Informationssuche	2 × 3 min
1b	Informationsauswahl anhand einer Liste mit bibliografischen Angaben	2 × 4 min
1c	Informationsauswahl anhand einer Liste mit Abstracts	2 × 7 min
2	Bewertung von Studien anhand von strukturierten Kurzfassungen	2 × 20 min (4 × 5 min)
3	evidenzbasierte Entscheidung und Argumentation	2 × 6 min

Tab. 1: Übersicht über Testteile und dafür vorgesehene Testzeiten

Informationsauswahl

Der Testteil zur *Informationsauswahl* umfasst bei jeder der beiden Entscheidungen zwei Teilaufgaben. Bei der *ersten Teilaufgabe* werden den Probanden bei jeder Entscheidung die zentralen bibliografischen Angaben (Autoren, Titel, Zeitschrift/Medium, Publikationsjahr) zu jeweils zehn Publikationen in einer Liste präsentiert, die in ihrem Aufbau den Trefferlisten sowohl üblicher digitaler Fachbibliografien (bspw. ERIC) als auch Suchmaschinen im Internet (z. B. Google Scholar) ähnelt (siehe Abb. 1). In den Listen sind jeweils vier hoch relevante, zwei eher relevante, zwei weniger relevante und zwei nicht relevante Publikationen enthalten. Die Probanden werden aufgefordert, die Studien auf einer Skala von 1 (auf keinen Fall) bis 4 (auf jeden Fall) im Hinblick darauf zu beurteilen, ob sie weitere Informationen darüber erhalten möchten. Die Ratings der Probanden werden mit einer im Projekt entwickelten Musterlösung abgeglichen. Wertet ein Proband eine hoch relevante Studie ebenfalls als hoch relevant (Antwort: auf jeden Fall), erhält er drei Punkte; weicht er mit seiner Einschätzung von der Musterlösung ab, erhält er für jeden Schritt Abstand einen Punkt weniger. Der maximal mögliche Wertebereich für ein Item reicht demnach von null bis drei Punkten. Der Gesamtwert für die erste Teilaufgabe der Informationsauswahl wird gebildet, indem die Bewertungen über die jeweils zehn Items bei den beiden Entscheidungen addiert werden. Auf dieser Skala können somit Werte bis 60 erzielt werden, wobei höhere Werte eine höhere Kompetenz in Bezug auf den ersten Teil der Informationsauswahl anzeigen.

Bei der *zweiten Teilaufgabe* werden den Probanden unabhängig von ihrer Auswahl beim vorhergehenden Schritt die Originalabstracts von sechs der zehn Publikationen, die bereits in der zuvor gezeigten Liste enthalten waren, vorgelegt (siehe Abb. 2). Vier dieser Publikationen sind für die vorliegende Entscheidung jeweils hoch relevant, zwei davon dagegen weniger relevant. Die Probanden werden daraufhin gebeten, jede der sechs Publikationen auf der Grundlage dieses erweiterten Informationsstandes auf einer Skala von 1 (stimme gar nicht zu) bis 7 (stimme völlig zu) im Hinblick darauf zu beurteilen, ob sie für die anstehende Entscheidung eine wichtige Rolle spielt. Die Ratings der Probanden werden einzeln mit der Musterlösung verglichen. Beurteilt ein Proband eine nicht relevante Studie ebenfalls als nicht relevant (Wert zwischen 1 und 3) oder eine relevante Studie als relevant (Wert zwischen 5 und 7), erhält er zwei Punkte; ent-



Einzelarbeit oder **Gruppenpuzzle**

Nach einigen Suchanfragen erhalten Sie die untenstehende Trefferliste. Bitte gehen Sie die Treffer durch und überlegen Sie, zu welcher der Studien Sie detailliertere Informationen haben möchten.

Wie schätzen Sie die Treffer ein?

Bitte wählen Sie eine der Antwortmöglichkeiten aus:

Auf Grundlage dieser Informationen möchte ich ...

... auf keinen Fall

... eher nicht

... eher schon

... auf jeden Fall

... mehr über die Studie wissen und eine Zusammenfassung ansehen.

Fördern Schulnoten die Motivation? Eine quasi-experimentelle Studie zum Einfluss der Benotungserwartung auf selbst berichtete und verhaltensnah erhobene Motivationsqualitäten

Hänze, M., Berger, R. & Bianchy, K - Psychologie in Erziehung und Unterricht - 2009

☐ auf keinen Fall ☐ eher nicht ☐ eher schon ☐ auf jeden Fall

Prozesse und Effekte „Kooperativen Lernens" im Sportunterricht

Bähr, I., Prohl, R. & Gröben, B. - Unterrichtswissenschaft - 2008

☐ auf keinen Fall ☐ eher nicht ☐ eher schon ☐ auf jeden Fall

Der Experteneffekt: Grenzen kooperativen Lernens in der Primarstufe?

Borsch, F., Gold, A., Kronenberger, J. & Souvignier, E. - Unterrichtswissenschaft - 2007

☐ auf keinen Fall ☐ eher nicht ☐ eher schon ☐ auf jeden Fall

Abb. 1: Screenshot der Teilaufgabe Informationsauswahl anhand einer Liste mit bibliografischen Angaben

scheidet er sich für die Mitte, erhält er einen Punkt. Schätzt der Proband eine Studie entgegen der Musterlösung ein, erhält er null Punkte. Der Gesamtwert für die zweite Teilaufgabe der Informationsauswahl wird gebildet, indem die Bewertungen über die jeweils sechs Items bei beiden Entscheidungen addiert werden. Die Skala umfasst somit Werte von 0 bis 24.

Beide Teilaufgaben der Informationsauswahl werden schließlich durch Addition zu einem Gesamtwert zusammengefasst. Insgesamt können für die Informationsauswahl somit maximal 84 Punkte erreicht werden, wobei höhere Werte eine höhere Kompetenz in Bezug auf die Informationsauswahl anzeigen.

Bewertung von Studien

In diesem Testteil werden den Probanden – wiederum unabhängig von ihrer vorherigen Auswahl – zu den vier relevanten unter den bereits bekannten Publikationen *strukturierte Kurzfassungen* vorgelegt (siehe Abb. 3). Dabei handelt es sich um ca. zweiseitige strukturierte Kurzfassungen wissenschaftlicher Originalveröffentlichungen mit standardisiertem Aufbau, der eng an die übliche Gliederung eines Berichts über eine empirische Untersuchung angelehnt ist. Jede dieser Studien wird von den Probanden im Hinblick

Abb. 2: Screenshot der Teilaufgabe Informationsauswahl anhand einer Liste mit Abstracts

auf die theoretisch abgeleiteten zehn Qualitätsaspekte beurteilt. Jeder dieser zehn Qualitätsaspekte wird auf einer Rating-Skala von 1 (stimme überhaupt nicht zu) bis 9 (stimme völlig zu) bewertet. Vor der Bearbeitung der beiden eigentlichen Entscheidungsaufgaben und nach Einführung des Szenarios zu Beginn der Testsitzung werden die Probanden anhand strukturierter Kurzfassungen von zwei weiteren Studien ohne Bezug zu den beiden Entscheidungen mit dieser Aufgabenstellung vertraut gemacht.

Das erste und das letzte Ratingitem (Bedeutung der Studie für die Entscheidung insgesamt ohne bzw. mit Berücksichtigung der einzelnen Qualitätsaspekte) werden bei der Bildung einer Skala für die *Bewertung von Studien* nicht berücksichtigt. Für jeden Probanden wird für jede der insgesamt acht Studien aus beiden Entscheidungen die Produkt-Moment-Korrelation zwischen den acht Ratings des Probanden und den acht Ratings aus einer im Projekt entwickelten Musterlösung berechnet. Der Wert pro Studie kann also ein Minimum von −1 und ein Maximum von +1 annehmen. Zur Bildung der Gesamtskala für die *Bewertung von Studien* wird der Mittelwert dieser acht Korrelationskoeffizienten gebildet, wobei höhere Werte eine höhere Kompetenz in Bezug auf die *Bewertung von Studien* anzeigen.

Einzelarbeit oder **Gruppenpuzzle**

Beurteilung von Untersuchungen

Bitte lesen Sie sich die im Folgenden angezeigte Kurzfassung einer wissenschaftlichen Untersuchung aufmerksam durch und geben Sie dann an, inwiefern sie den unten stehenden Aussagen zustimmen.

Dafür sind insgesamt etwa fünf Minuten Bearbeitungszeit vorgesehen.

Das Gruppenpuzzle im Physikunterricht der Sekundarstufe II – Einfluss auf Motivation, Lernen und Leistung

M. Hänze und R. Berger

In dieser Studie wurde untersucht, ob die grundlegenden Bedürfnisse nach sozialer Eingebundenheit, Kompetenz- und Autonomieerleben beim Lernen nach der Methode des Gruppenpuzzles in höherem Ausmaß befriedigt werden als im Frontalunterricht, und ob infolgedessen ein höherer Lernerfolg erzielt wird.

Methodik

Teilnehmer. In die Studie wurden Daten von 61 Zwölftklässlern aus acht Physik-Grundkursen an fünf Schulen einbezogen, die bei zwei Unterrichtseinheiten zu den Themen „Rasterelektronenmikroskop“ und „Mikrowellenofen“ in zwei verschiedenen Halbjahren anwesend waren.

Untersuchungsplan. Bei jeder Unterrichtseinheit wurde ein Teil der Grundkurse nach der Methode des Gruppenpuzzles unterrichtet, im anderen Teil fand Frontalunterricht statt. Bei der zweiten Unterrichtseinheit wurde die Zuordnung der Kurse zu den Bedingungen umgekehrt.

Unterrichtseinheiten und Ablauf. Die Unterrichtseinheit zum Thema „Rasterelektronenmikroskop“ umfasste vier Schulstunden, die zum Thema „Mikrowellenofen“ drei Schulstunden. Vor jeder Unterrichtseinheit wurde das Vorwissen zum Thema erfasst. Im Anschluss wurde jeweils ein unbenoteter und unangekündigter Abschlusstest durchgeführt.

	Ich stimme überhaupt nicht zu								**Ich stimme völlig zu**
Die Studie spielt für die aktuelle Entscheidung eine wichtige Rolle.	○	○	○	○	○	○	○	○	○
Aus dem Text geht hervor, dass die für die Entscheidung relevante didaktische Gestaltungsvariante untersucht wurde.	○	○	○	○	○	○	○	○	○
Aus dem Text geht hervor, dass die für dieEntscheidung relevante didaktische Gestaltungsvariante im Kontext der aktuellen Entscheidung leicht umsetzbar ist.	○	○	○	○	○	○	○	○	○
Aus dem Text geht hervor, dass die untersuchten Erfolgskriterien mit den Erfolgskriterien bei der aktuellen Entscheidung überein stimmen.	○	○	○	○	○	○	○	○	○
Aus dem Text geht hervor, dass die Erfolgskriterien mit geeigneten Verfahren gemessen wurden.	○	○	○	○	○	○	○	○	○
Aus dem Text geht hervor, dass die Rahmenbedingungen mit den Rahmenbedingungen bei der aktuellen Entscheidung in den wichtigen Merkmalen übereinstimmen.	○	○	○	○	○	○	○	○	○
Aus dem Text geht hervor, dass die Teilnehmer mit der Zielgruppe bei der aktuellen Entscheidung in den wichtigen Merkmalen übereinstimmen.	○	○	○	○	○	○	○	○	○
Aus dem Text geht hervor, dass die für die Entscheidung relevanten Ergebnisse statistisch belastbar sind.	○	○	○	○	○	○	○	○	○
Soweit ersichtlich, sind die für die Entscheidung relevanten Effekte eindeutig auf die untersuchte Gestaltungsvariante zurückzuführen.	○	○	○	○	○	○	○	○	○
Unter Berücksichtigung der zuvor bewerteten Aspekte spielt die Studie für die aktuelle Entscheidung eine wichtige Rolle.	○	○	○	○	○	○	○	○	○

Abb. 3: Screenshot der Teilaufgabe Bewertung von Studien

Hintergrundvariablen

Allgemeine kognitive Fähigkeiten. Diese wurden zum einen mit einer selbst zusammengestellten Kurzfassung der verbalen Analogien aus dem *Intelligenz-Struktur-Test* (I-S-T 2000R; Liepmann, Beauducal, Brocke & Amthauer, 2007) sowie einer validierten Kurzfassung des *Raven's Advanced Progressive Matrices Test* (Bors & Stokes, 1998; Raven & Court, 1988) erfasst.

Wissen. Es wurden zwei Wissensbereiche berücksichtigt: Der Fragebogen zum *pädagogischen Wissen* wurde im Projekt entwickelt und umfasst Fragen zu Theorien des Lehrens und Lernens mit Aufgaben im Single-Choice-Format. Ein Beispielitem lautet: „Welches der folgenden Unterrichtsmerkmale gehört nicht zu den Qualitätsmerkmalen guten Unterrichts nach Helmke?" (richtige Antwort: Präzision; Distraktoren: Klarheit und Strukturiertheit, Aktivierung, Konsolidierung und Sicherung). Die Skala *forschungsmethodisches Wissen* wurde ebenfalls selbst entwickelt und erfasst das Wissen der Probanden in Hinblick auf methodische Merkmale wissenschaftlicher Studien mit Aufgaben im Single-Choice-Format. Ein Beispielitem lautet: „Was versteht man unter „Reliabilität"? (richtige Antwort: die Messgenauigkeit eines Messinstruments; Distraktoren: die Aussagekraft eines Tests für das zu erfassende Merkmal, das Ausmaß, in dem Testergebnisse von den durchführenden Personen unabhängig sind, die Belastbarkeit der Ergebnisse einer Untersuchung).

Wissenschaftliches Denken (scientific reasoning). Die Skala erfasst die Fähigkeit der Probanden, wissenschaftliche Schlussfolgerungen zu ziehen. Für die Erfassung dieser Fähigkeit wurden Items im Single-Choice-Format aus dem Instrument von Lippman (2012) verwendet.

Kritisches Denken (critical thinking). Die Skala umfasst die Fähigkeit des induktiven Schlussfolgerns und Reflektierens von Evidenz. Die Aufgaben im Single-Choice-Format wurden ebenfalls von Lippman (2012) übernommen. Die Skalen zum wissenschaftlichen und kritischen Denken wurden jeweils durch die Berechnung der Summe der Punkte bei den Einzelitems gebildet.

Epistemologische Überzeugungen. Diese wurden mit einer deutschen Fassung des *Discipline-Focused Epistemological Beliefs Questionnaire* (Hofer, 2000) auf einer Skala von 1 (stimme nicht zu) bis 5 (stimme zu) erfasst. Zur Berechnung des Gesamtscores wurde ein Mittelwert gebildet.

Überzeugungen zu forschungsmethodischem Wissen. Diese Skala erfasst Selbsteinschätzungen zum Verständnis forschungsbezogener Fachbegriffe (adaptiert nach Bauer & Prenzel, 2012; Jette et al., 2003) auf einer Skala von 1 (ich kenne den Begriff nicht) bis 4 (ich verstehe den Begriff und könnte ihn anderen erklären). Ein Beispielitem lautet: „Wenn Sie einen bildungswissenschaftlichen Forschungsbericht lesen, könnten darin unter anderem folgende Fachbegriffe auftauchen: a. Evidenz, b. Randomisiertes Kontrollgruppendesign, c. Quasi-Experiment, d. Meta-Analyse, e. Repräsentative Stichprobe, f. Effektstärke. Wie gut schätzen Sie Ihr Verständnis dieser Fachbegriffe ein?". Die statistischen Kennwerte der Variablen im Vortest sind in Tabelle 2 aufgelistet.

Variablen	Anzahl Items	M	SD	Cronbachs α	Min	Max
Allgemeine kognitive Fähigkeiten – verbale Analogien	10	5.86	1.90	.53	0	10
Allgemeine kognitive Fähigkeiten – figurale Analogien	12	6.69	2.54	.71	0	12
Forschungsmethodisches Wissen	16	3.53	1.65	.46	0	8
Pädagogisches Wissen	10	5.07	1.72	.31	0	10
Wissenschaftliches Denken	10	4.95	2.43	.65	0	10
Kritisches Denken	15	8.11	2.34	.35	0	13
Epistemologische Überzeugungen	18	2.32	.39	.69	–	–
Überzeugungen zu forschungsmethodischem Wissen	10	2.96	.58	.78	–	–

Tab. 2: Skalenkennwerte der Hintergrundvariablen

3.2 Ablauf der Untersuchung

Vor der Durchführung des Kompetenztests wurden alle Probanden gebeten, einen Online-Test für die Hintergrundvariablen (siehe Tab. 2) sowie demografische Angaben auszufüllen. Der entsprechende Link zum Online-Test wurde per Mail an die Probanden versendet. Die Bearbeitung des Kompetenztests wurde mit allen Studierenden in Präsenz mit je einer Versuchsleiterin bzw. einem Versuchsleiter unter Verwendung eines standardisierten Leitfadens durchgeführt. Die mittlere Bearbeitungszeit betrug 74 Minuten für den Haupttest und 57 Minuten für den Vortest.

3.3 Stichprobe

Die folgenden Analysen beruhen auf einer Stichprobe von insgesamt 341 Studierenden. Von diesen waren 78.8 % weiblich und 17.4 % männlich (3.8 % fehlende Angaben). Das durchschnittliche Alter lag bei $M = 23.13$ Jahren ($SD = 5.27$). Unter den Teilnehmern befanden sich 157 Masterstudierende in den Bildungswissenschaften, 109 Studierende eines Bachelorstudiengangs Pädagogik/Bildungswissenschaft, 36 Lehramtsstudierende und 39 Promotionsstudierende aus den Bereichen Bildungswissenschaften, Pädagogik und Psychologie. Die Master-, Bachelor- und Lehramtsstudierenden stammten von der Bergischen Universität Wuppertal und der Ludwig-Maximilians-Universität München. Die befragten Promotionsstudierenden wurden deutschlandweit rekrutiert und absolvierten die Testsitzung online. Damit ist die vergleichsweise geringe Stichprobengröße in dieser Gruppe zu erklären. Der Großteil der Teilnehmer (86.9 %) hat Deutsch als Muttersprache.

4. Ergebnisse

4.1 Inhaltsvalidität

Informationsauswahl
Die inhaltliche Validierung der Teilkompetenz beruht auf der Befragung von Experten. Hierfür wurden die konstruierten Trefferlisten in zufälliger Reihenfolge zusammen mit den nötigen Hintergrundinformationen aus dem Szenario elf wissenschaftlichen Mitarbeitern (Doktoranden, Postdoktoranden, Professoren) am Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie der Universität München zur Bearbeitung entsprechend der Testaufgabe vorgelegt. Als Kennwert für die Übereinstimmung der Beurteilungen wurde eine Intraklassenkorrelation über alle Beurteiler und über beide Entscheidungen hinweg ermittelt ($ICC_{just.MW} = .95$).

Bewertung von Studien
Für die inhaltliche Validierung der Teilkompetenz *Bewertung von Studien* wurden Experten in verschiedener Weise eingebunden: Zunächst wurde die Teilaufgabe von zwei wissenschaftlichen Mitarbeitern der LMU München mit einschlägigen Kenntnissen im Bereich der Evidenzbasierung bearbeitet. Die initiale Beurteilerübereinstimmung wurde anhand einer Intraklassenkorrelation berechnet ($ICC_{just.MW} = .78$). Die Lösungen der Experten wurden verglichen, und differierende Bewertungen wurden konsensuell ausdiskutiert. Diese Musterlösung wurde mit einem weiteren Projektmitarbeiter und Experten diskutiert und nach Diskussion der Hauptpunkte noch einmal angeglichen.

4.2 Konstruktvalidität

In Bezug auf die Konstruktvalidierung wurde die interne Konsistenz (Cronbachs α) der beiden Teilkompetenzen *Informationsauswahl* und *Bewertung von Studien* sowie der Zusammenhang dieser beiden Teilskalen ermittelt. Die interne Konsistenz der Teilkompetenz der *Informationsauswahl* lag bei Cronbachs $\alpha = .65$, die der Teilkompetenz *Bewertung von Studien* bei Cronbachs $\alpha = .53$. Die beiden Teilkompetenzen korrelierten signifikant miteinander ($r = .22$, $p < .01$). Zudem wurden Korrelationen zwischen den einbezogenen Teilkompetenzen *Informationsauswahl* und *Bewertung von Studien* und den erhobenen Hintergrundvariablen berechnet. Die Ergebnisse (siehe Tab. 3) zeigen signifikante Korrelationen der Teilkompetenz *Informationsauswahl* mit den verbalen Analogien aus dem Intelligenz-Struktur-Test, der Fähigkeit des kritischen Denkens, den Überzeugungen zu forschungsmethodischem Wissen, dem pädagogischen Wissen und der Fähigkeit des wissenschaftlichen Denkens. Die Teilkompetenz *Bewertung von Studien* weist einen signifikanten Zusammenhang mit allen einbezogenen Variablen auf. Insbesondere die verbalen Analogien aus dem Intelligenz-Struktur-Test, die epistemologischen Überzeugungen und das wissenschaftliche Denken weisen die höchsten positiven Zusammenhänge mit dieser Teilkompetenz auf.

	Informationsauswahl	Bewertung von Studien
Allgemeine kognitive Fähigkeiten		
Verbale Analogien	.21**	.25**
Figurale Analogien	.12	.15*
Forschungsmethodisches Wissen	.04	.16**
Pädagogisches Wissen	.13*	.19**
Wissenschaftliches Denken	.15*	.23**
Kritisches Denken	.17**	.15*
Überzeugungen zu forschungsmethodischem Wissen	.18**	.17**
Epistemologische Überzeugungen	.08	.24**

** $p < .01$; * $p < .05$

Tab. 3: Korrelation der Teilkompetenzen Informationsauswahl und Bewertung von Studien mit den Hintergrundvariablen

Um gemeinsame Varianzanteile der Hintergrundvariablen angemessen zu berücksichtigen, wurden anschließend multiple Regressionsanalysen mit den beiden Teilkompetenzen als abhängigen Variablen und den übrigen Variablen als Prädiktoren durchgeführt (siehe Tab. 4). Die Hintergrundvariablen wurden dafür in vier theoretisch gebildeten Blöcken (hierarchisch) eingeschlossen, um den individuellen Varianzanteil der einzelnen Blöcke am Gesamtmodell zu erhalten: Der erste Block bildet die allgemeinen kognitiven Fähigkeiten ab, der zweite Block die Wissensvariablen und der dritte Block die Fähigkeiten zum wissenschaftlichen und kritischen Denken. Im vierten Block wurden die übrigen Variablen Überzeugungen zu forschungsmethodischem Wissen sowie epistemologische Überzeugungen zusammengefasst. Für die Teilkompetenz *Informationsauswahl* zeigt sich, dass allgemeine kognitive Fähigkeiten fünf Prozent der Varianz aufklären. Weitere Prädiktoren spielen für die Vorhersage der Informationsauswahl keine statistisch belastbare Rolle. Eine Regressionsanalyse mit allen Prädiktoren klärt einen Varianzanteil von acht Prozent auf ($F(8, 232) = 2.69$, $p < .01$).

Bei der Teilkompetenz *Bewertung von Studien* zeigt sich ein anderes Bild: Allgemeine kognitive Fähigkeiten klären einen signifikanten Varianzanteil von sieben Prozent auf. Forschungsmethodisches und pädagogisches Wissen klären gemeinsam weitere zwei Prozent der Varianz auf; dieser Zuwachs an aufgeklärtem Varianzanteil ist jedoch nicht signifikant. Die Fähigkeiten des wissenschaftlichen und des kritischen Denkens erklären gemeinsam signifikant weitere zwei Prozent der Varianz der Teilkompetenz *Bewertung von Studien.* Epistemologische Überzeugungen leisten einen signi-

	Informationsauswahl				Bewertung von Studien			
Prädiktoren	**Modell 1**	**Modell 2**	**Modell 3**	**Modell 4**	**Modell 1**	**Modell 2**	**Modell 3**	**Modell 4**
Allgemeine kognitive Fähigkeiten:								
Verbale Analogien	.19**	.19*	.18*	.17*	.26**	.19**	.16*	.17*
Figurale Analogien	.34	.02	−.02	−.01	.06	.04	−.02	−.04
Forschungsmethodisches Wissen		−.03	−.01	−.04		.09	.09	.09
Pädagogisches Wissen		.08	.07	.04		.10	.08	.03
Wissenschaftliches Denken			.04	.02			.12	.11
Kritisches Denken			.13(*)	.11			.10	.09
Überzeugungen zu forschungsmethodischem Wissen				.13				.04
Epistemologische Überzeugungen				.01				.17**
Änderung in R^2	.05**	.01	.02	.01	.07**	.02	.02*	.03*

** $p < .01$; * $p < .05$; (*) $p < .07$

Tab. 4: Standardisierte Regressionskoeffizienten aus multiplen hierarchischen Regressionen mit den Teilkompetenzen als abhängige und den Hintergrundvariablen als unabhängige Variablen

fikanten Beitrag zur Vorhersage der Teilkompetenz *Bewertung von Studien* und klären gemeinsam mit den Überzeugungen zu forschungsmethodischem Wissen weitere drei Prozent der Varianz auf. Insgesamt können durch die Regressionsgleichung mit allen übrigen Variablen als Prädiktoren 14 Prozent der Varianz der Teilkompetenz *Bewertung von Studien* erklärt werden ($F(8, 231) = 4.54$, $p < .01$).

5. Diskussion

Die Umsetzung einer stärker evidenzbasierten Praxis im Bildungsbereich auf der Ebene von individuellen Entscheidungen stellt hohe kognitive Anforderungen an pädagogische Praktiker (Bromme et al., 2014, S. 8). Ein ganzes Bündel von Fertigkeiten muss hierfür eingesetzt werden. Zentral sind die Bereiche der *Informationsauswahl* und der *Bewertung von Studien.* Für diese Teilkompetenzen der Kompetenz im evidenzbasierten Argumentieren wurde im vorliegenden Beitrag anhand von kognitionspsychologischen Erkenntnissen ein theoretisches Modell entwickelt, und anschließend wurden empirische Argumente für die Validität der entwickelten Instrumente gesammelt. Die zentralen Ergebnisse dieser Analysen lauten:

1) Die Teilkompetenzen *Informationsauswahl* und *Bewertung von Studien* sind empirisch voneinander unterscheidbar. Der schwache, wenngleich signifikante Zusammenhang zwischen den beiden Teilkompetenzen belegt empirisch das theoretisch erwartete Befundmuster der voneinander trennbaren Teilkompetenzen *Informationsauswahl* und *Bewertung von Studien.* Diese beruhen auf hinreichend unterschiedlichen kognitiven Aktivitäten. Zudem weisen sie einen erwartbaren Zusammenhang mit Maßen für allgemeine kognitive Fähigkeiten auf, sind aber dennoch deutlich von diesen unterscheidbar. Dieses Ergebnis kann als Indikator für die Konstruktvalidität gewertet werden (Cronbach & Meehl, 1955).
2) Betrachtet man die Korrelationen zwischen den beiden untersuchten Teilkompetenzen und den Hintergrundvariablen, bietet sich folgendes Bild: Beide Teilkompetenzen hängen mit allgemeinen (insbesondere verbalen) kognitiven Fähigkeiten zusammen. Bei der *Informationsauswahl* scheint ein Zusammenhang mit dem pädagogischen Wissen und bei der *Bewertung von Studien* zusätzlich ein Zusammenhang mit dem forschungsmethodischen Wissen zu bestehen. Diese Befunde wären in Anbetracht der unterschiedlichen Anforderungen bei den beiden Teilkompetenzen plausibel und theoretisch erwartbar (Brand-Gruwel et al., 2005): Für die *Informationsauswahl* dürfte pädagogisches Wissen wichtig sein – z. B. für das Verständnis von pädagogischen Fachbegriffen. Bei der *Bewertung von Studien* könnte darüber hinaus forschungsmethodischem Wissen eine wichtige Rolle zukommen: Je höher das forschungsmethodische Wissen ausgeprägt ist, desto besser dürfte eine Bewertung von Studien einschließlich ihrer forschungsmethodischen Qualität vorgenommen werden können. Diese Zusammenhänge mit dem pädagogischen und dem forschungsmethodischen Wissen verschwinden jedoch allesamt, wenn die allgemeinen kognitiven Fähigkeiten statistisch kontrolliert werden. Möglicherweise leistet Wissen bei der Anwendung der beiden Teilkompetenzen keinen Beitrag über das Ausmaß hinaus, in dem es selbst Niederschlag allgemeiner kognitiver Fähigkeiten ist. Derartige Erklärungsansätze erfordern jedoch weitere empirische Untersuchungen. Ähnliches gilt für die Rolle des wissenschaftlichen und des kritischen Denkens.

Zusammenfassend wollen wir folgende Punkte noch einmal festhalten: In diesem Beitrag konnten empirische Hinweise dafür gesammelt werden, dass die Kompetenz im evidenzbasierten Argumentieren aus wenigstens zwei zu differenzierenden Teilkompetenzen besteht. Die entwickelten Testinstrumente stellen eine Weiterentwicklung des zugrunde liegenden Modells des *Information Problem Solving* dar (Lazonder & Rouet, 2008), in dem die Teilkompetenz der *Informationsauswahl* um die Teilkompetenz zur systematischen *Bewertung von Studien* ergänzt wurde. Das Bewältigen der komplexen Anforderungen bei der *Informationsauswahl* und der *Bewertung von Studien* kann nicht durch anforderungsspezifisches Wissen, allgemeine kognitive Fähigkeiten und epistemologische Überzeugungen allein erklärt werden. Es ist plausibel anzunehmen, dass – das zeigen die Erfahrungen im Bereich der Medizin – die komplexe Kompetenz, evidenzbasiert zu argumentieren, gefördert werden kann. Dies auch für den Bildungsbereich nachzuweisen, sollte Gegenstand zukünftiger Forschung sein. Mit Blick auf die

Standards für Lehrerbildung der KMK hat die vorliegende Studie Hinweise geliefert, dass die untersuchten Teilkompetenzen differenziert betrachtet werden sollten (KMK, 2014). Außerdem legen die gewonnenen Erkenntnisse nahe, dass noch weitere Teilkompetenzen der Kompetenz im evidenzbasierten Argumentieren theoretisch ausdifferenziert und empirisch bestätigt werden sollten (siehe Tab. 1, z. B. die Rechtfertigung einer evidenzbasierten Entscheidung).

Nächste Schritte zur Weiterentwicklung der evidenzbasierten Praxis im Bildungsbereich könnten durch die Abstimmung von Studieninhalten und die Integration entsprechender Inhalte in Maßnahmen zur Qualitätssicherung und Weiterbildung von pädagogischen Fachkräften angegangen werden. Weitere Stärkung könnte die Idee evidenzbasierter pädagogischer Praxis erfahren, wenn es zukünftiger Forschung gelänge, die ganze Wirkungskette von der Planung bis zur Umsetzung in der Lehre und den Resultaten dieser Praxis bei den Lernenden in den Blick zu nehmen.

Literatur

Amelang, M. (2000). Intelligenz. In M. Amelang (Hrsg.), *Enzyklopädie der Psychologie. Differentielle Psychologie und Persönlichkeitsforschung. Bd. 4: Determinanten individueller Unterschiede* (S. 245–328). Göttingen: Hogrefe.

Astleitner, H., Brünken, R., & Zander, S. (2002). Können Schüler und Lehrer kritisch denken? Lösungserfolg und -strategien bei typischen Aufgaben. *Salzburger Beiträge zur Erziehungswissenschaft, 6*(2), 51–61.

Bauer, J., & Prenzel, M. (2012). *Scientific and professional journals as resource for teachers' professional learning and evidence-based practice*. Vortrag auf der EARLI SIG 14 (Learning and Professional Development) Conference „Learning in transition", Antwerpen, 22. bis 24. August 2012.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K., & Weiß, M. (Hrsg.) (2002). *PISA 2000: Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*(4), 469–520.

Beelmann, A. (2014). Möglichkeiten und Grenzen systematischer Evidenzkumulation durch Forschungssynthesen in der Bildungsforschung. *Zeitschrift für Erziehungswissenschaft, 17*(4), 55–78.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*(3), 382–398.

Brand-Gruwel, S., Wopereis, I. G. J. H., & Vermetten, Y. (2005). Information problem solving by experts and novices: Analysis of a complex cognitive skill. *Computers in Human Behavior, 21*, 487–508.

Brand-Gruwel, S., Wopereis, I. G. J. H., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education, 53*(4), 1207–1217.

Bromme, R. (2014). *Der Lehrer als Experte: Zur Psychologie des professionellen Wissens* (2. Aufl.). Münster: Waxmann.

Bromme, R., Prenzel, M., & Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Zeitschrift für Erziehungswissenschaft, 17*(4), 3–54.

Brown, N. J. S., Furtak, E. M., Timms, M., Nagashima, S. O., & Wilson, M. (2010). The Evidence-Based Reasoning Framework: Assessing scientific reasoning. *Educational Assessment, 15*(3/4), 123–141.

Cook, B. G., Smith, G. J., & Tankersley, M. (2012). Evidence-based practices in education. In K. R. Harris, S. Graham & T. Urdan (Hrsg.), *APA Educational Psychology Handbook. Bd. 1: Theories, constructs, and critical issues* (S. 495–528). Washingon, D. C.: American Psychological Association.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Cook, T. D., & Gorard, S. (2007). What counts and what should count as evidence. In OECD (Hrsg.), *Evidence in education: Linking research and policy* (S. 33–49). Paris: OECD.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programmes*. San Francisco: Jossey-Bass.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Evidence-Based Medicine Working Group (EBMWG) (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268*(17), 2420–2425.

Feltovich, P. J., Prietula, M. J., & Ericsson, K. A. (2006). Studies of expertise from psychological perspectives. In K. A. Ericcson (Hrsg.), *The Cambridge handbook of expertise and expert performance* (S. 41–67). New York: Cambridge University Press.

Fischer, F., Waibel, M., & Wecker, C. (2005). Nutzenorientierte Grundlagenforschung im Bildungsbereich. *Zeitschrift für Erziehungswissenschaft, 8*(3), 427–442.

Frey, A., & Hartig, J. (2013). Wann sollten computerbasierte Verfahren zur Messung von Kompetenzen anstelle von papier- und bleistift-basierten Verfahren eingesetzt werden? *Zeitschrift für Erziehungswissenschaft, 16*(1), 53–57.

Goldman, S. R., Lawless, K. A., Pellegrino, J., Braasch, J., & Gomez, K. (2012). A technology for assessing multiple source comprehension: An essential skill of the 21st century. In M. C. Mayrath, J. Clarke-Midura & D. H. Robinson (Hrsg.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (S. 173–209). Charlotte: Information Age Publishing.

Harden, M., Grant, J., Buckley, G., & Hart, I. R. (1999). BEME Guide No. 1: Best evidence medical education. *Medical Teacher, 21*(6), 553–62.

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 143–172). Wiesbaden: Springer.

Hetmanek, A., Wecker, C., Trempler, K., Kiesewetter, J., Hake, C., Gräsel, C., Fischer, M. R., & Fischer, F. (2014). *Lowering the bar for evidence-based practice in teaching – what structured abstracts may contribute*. Vortrag auf der EARLI SIG 11 (Teaching and Teacher Education) Conference „Practice-Oriented Teacher Learning and Professional Development“, Frauenchiemsee, 16. bis 18. Juni 2014.

Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemporary Educational Pychology, 25*, 378–405.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67*(1), 88–140.

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik, 61. Beiheft*, 11–31.

Jette, D. U., Bacon, K., Batty, C., Carlson, M., Ferland, A., Hemingway, R. D., Hill, J., Ogilvie, L., & Volk, D. (2003). Evidence-based practice: Beliefs, attitudes, knowledge, and behaviors of physical therapists. *Journal of the American Physical Therapy, 83*, 786–805.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Klieme, E., Funke, J., Leutner, D., Reimann, P., & Wirth, J. (2001). Problemlösen als fächerübergreifende Kompetenz. Konzeption und erste Resultate aus einer Schulleistungsstudie. *Zeitschrift für Pädagogik, 47*(2), 179–200.

Ständige Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (2014). *Standards für die Lehrerbildung: Bildungswissenschaften*. http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf [05. 08. 2014].

Lazonder, A. W., & Rouet, J.-F. (2008). Information problem solving instruction: Some cognitive and metacognitive issues. *Computers in Human Behavior, 24*(3), 753–765.

Liepmann, D., Beauducal, A., Brocke, B., & Amthauer, R. (2007). *I-S-T 2000 R* (2. Aufl.). Göttingen: Hogrefe.

Lippman, J. P. (2012). *Improving and predicting novice reasoning about the evidentiary connection between psychological studies and theories* (unveröffentlichte Dissertation). Chicago: University of Illinois.

Mayer, R. E. (2003). What causes individual differences in cognitive performance. In R. J. Sternber & E. L. Grigorenko (Hrsg.), *The psychology of abilities, competencies, and expertise* (S. 263–273). Cambridge: Cambridge University Press.

Montori, V., & Guyatt, G. (2008). Progress in evidence-based medicine. *Journal of the American Medical Association, 300*(15), 1814–1816.

Pant, H. A. (2014). Aufbereitung von Evidenz für bildungspolitische und pädagogische Entscheidungen: Metaanalysen in der Bildungsforschung. *Zeitschrift für Erziehungswissenschaft, 17*(4), 79–99.

Raven, J. C., & Court, J. H. (1988). *Raven's Progressive Matrices und Vocabulary Scales. Manual. Teil 4: Advanced Progressive Matrices*. Set I & II. Frankfurt: Swets & Zeitlinger.

Rosenberg, W. M. C., & Donald, A. (1995). Evidence based medicine: An approach to clinical problem-solving. *BMJ: British Medical Journal, 310*(6987), 1122–1126.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, S. W. (1996). Evidence based medicine: What it is and what it isn't. *BMJ: British Medical Journal, 312* (7023), 71–72.

Simon, H. A. (1973). The structure of ill-structured problems. *Artificial Intelligence, 4*, 181–201.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5–14.

Tugwell, P., Haynes, R. B., & Sackett, D. L. (1992). *Clinical epidemiology: A basic science for clinical medicine.* Boston: Little, Brown and Company.

Wecker, C. (2013). How to support prescriptive statements by empirical research: Some missing parts. *Educational Psychology Review, 25*(1), 1–18.

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal, 46*(4), 1060–1106.

Wopereis, I. G. J. H., & van Merriënboer, J. J. G. (2011). Evaluating text-based information on the World Wide Web. *Learning and Instruction, 21*(2), 232–237.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*(1), 99–149.

Abstract: Following the model of medicine, evidence-based practice can be characterized as the conscientious use of the best currently available empirical evidence in professional practice. This is a complex challenge for educational practitioners. The goal of the present paper is to validate an instrument for the measurement of the required competence. The focus is on the two sub-competences of *information selection* and *evaluation of studies.* The participants were 341 students from degree programs in education. They completed a case-based online test for information selection and evaluation of studies. The results indicate that the two sub-competences can be separated and exhibit plausible correlations to other variables.

Keywords: Validation, Competency Assessment, Case-Based Decision, Evidence-Based Practice in Education, Empirical Studies

Anschrift der Autor(inn)en

Kati Trempler, Bergische Universität Wuppertal, School of Education,
Institut für Bildungsforschung, Gaußstraße 20, 42119 Wuppertal, Deutschland
E-Mail: trempler@uni-wuppertal.de

M.A. Andreas Hetmanek, Ludwig-Maximilians Universität München,
Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie,
Leopoldstraße 13, 80803 München, Deutschland
E-Mail: andreas.hetmanek@psy.lmu.de

PD Dr. Christof Wecker, Ludwig-Maximilians Universität München,
Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie,
Leopoldstraße 13, 80803 München, Deutschland
E-Mail: christof.wecker@psy.lmu.de

Dr. Jan Kiesewetter, Institut für Didaktik und Ausbildungsforschung in der Medizin,
Klinikum der Universität München, Ziemssenstraße 1, 80336 München, Deutschland
E-Mail: jan.kiesewetter@med.uni-muenchen.de

Mia Wermelt, Institut für Didaktik und Ausbildungsforschung in der Medizin,
Klinikum der Universität München, Ziemssenstraße 1, 80336 München, Deutschland
E-Mail: mia.wermelt@med.uni-muenchen.de

Prof. Dr. Frank Fischer, Ludwig-Maximilians Universität München,
Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie,
Leopoldstraße 13, 80803 München, Deutschland
E-Mail: Frank.Fischer@psy.lmu.de

Prof. Dr. Martin Fischer, Institut für Didaktik und Ausbildungsforschung in der Medizin,
Klinikum der Universität München, Ziemssenstraße 1, 80336 München, Deutschland
E-Mail: martin.fischer@med.uni-muenchen.de

Prof. Dr. Cornelia Gräsel, Bergische Universität Wuppertal, School of Education,
Institut für Bildungsforschung, Gaußstraße 20, 42119 Wuppertal, Deutschland
E-Mail: graesel@uni-wuppertal.de

Sandra Schladitz/Jana Groß Ophoff/Markus Wirtz

Konstruktvalidierung eines Tests zur Messung bildungswissenschaftlicher Forschungskompetenz

Zusammenfassung: Die Befähigung, auf wissenschaftlicher Evidenz basierende Entscheidungen zu treffen, ist zentrales Ziel einer Hochschulausbildung. Das Projekt LeScEd (Learning the Science of Education) integriert Ansätze aus Bereichen wie Informationswissenschaften, Mathematikdidaktik und Psychologie in ein gemeinsames Strukturmodell bildungswissenschaftlicher Forschungskompetenz (BFK). Die aktuelle Studie untersucht Zusammenhänge der BFK zu fluider Intelligenz und selbsteingeschätzter Kompetenz im Sinne diskriminanter bzw. konvergenter Validierung. In Strukturgleichungsmodellen zeigen sich kleine positive Zusammenhänge zur Intelligenz, aber keine zur Selbsteinschätzung. Dies deutet darauf hin, dass BFK mit Intelligenz verwandt, aber von ihr abgrenzbar ist und dass die Selbsteinschätzung keinen geeigneten Indikator für die tatsächliche Kompetenz darstellt.

Schlagworte: bildungswissenschaftliche Forschungskompetenz, Hochschulforschung, Intelligenz, Konstruktvalidität, Strukturgleichungsmodelle

Evidenzbasierte Entscheidungen treffen zu können, ist ein zentrales Qualifizierungsziel eines Hochschulstudiums (KMK, 2005). Voraussetzung hierfür ist unter anderem die Kompetenz, den Forschungsstand des jeweiligen Fachgebietes auf Basis der Literatur strukturieren, verstehen sowie problembezogen evaluieren und reflektieren zu können. Im Bereich der Bildungswissenschaften wird dies als *Educational Research Literacy* (Shank & Brown, 2007; McMillan & Schumacher, 2010) bzw. als bildungswissenschaftliche Forschungskompetenz bezeichnet (BFK; Groß Ophoff, Schladitz, Lohrmann & Wirtz, im Druck). Im Rahmen der evidenzbasierten Beschreibung und Messung von Kompetenzen ist eine differenzierte Konstruktvalidierung insbesondere hinsichtlich konvergenter und diskriminanter Validierung anhand inhaltlich relevanter Referenzkonstrukte erforderlich. Nach Weinert (2001) sollten objektiv erfasste Kompetenzen von u. a. fluider Intelligenz (als domänenübergreifende und nicht veränderbare kognitive Grundfähigkeit) sowie einer subjektiven Kompetenzeinschätzung (als Überzeugung bzgl. der eigenen Fähigkeiten) psychometrisch abgegrenzt werden. Diese Differenzierungen werden im Folgenden auf die BFK übertragen.

1. Theoretischer Hintergrund

Die zunehmende Tendenz zur Bilanzierung von Bildungsprozessen hat sich in den vergangenen Jahren in groß angelegten Schulleistungsstudien wie beispielsweise PISA (Programme for International Student Assessment; Klieme et al., 2010) gezeigt. Die Output-Orientierung gewinnt auch im deutschen Bildungswesen im Hochschulbereich zunehmend an Bedeutung (Zlatkin-Troitschanskaia & Kuhn, 2010). Um Bildungsprozesse evaluieren zu können, ist zunächst ein präzises Verständnis von zugrunde liegenden Kompetenzstrukturen erforderlich (Klieme & Leutner, 2006). Entsprechend wurden in den vergangenen Jahren zahlreiche Projekte ins Leben gerufen, die sich mit der Untersuchung von Kompetenzen im tertiären Bildungsbereich auseinandersetzen (z. B. Teacher Education and Development Study: Mathematics, TEDS-M; Blömeke, Kaiser & Lehmann, 2010). Dabei standen immer häufiger auch Studierende der Bildungswissenschaften, speziell des Lehramts, im Mittelpunkt des Interesses, da diese zukünftigen Generationen kompetent Wissen vermitteln und sie adäquat fördern sollen (Terhart, 2012). AbsolventInnen müssen nicht nur über fachliche und pädagogische Kompetenzen verfügen, sondern auch in der Lage sein, in ihrer pädagogischen Praxis evidenzbasiert Entscheidungen treffen und sich weiterbilden zu können (KMK, 2004). Voraussetzung dafür ist das Grundverständnis, dass zur Beantwortung von Fragestellungen bzw. zur Lösung von Problemen der Bezug auf fundierte wissenschaftliche Erkenntnisse notwendig ist (Brown, Furtak, Timms, Nagashima & Wilson, 2010). Darauf aufbauend ist es wichtig, zielgerichtet Forschungsergebnisse und wissenschaftliche Literatur erschließen, diese kritisch reflektieren sowie daraus handlungsleitende Schlussfolgerungen ableiten zu können (Shank & Brown, 2007).

In der Literatur lassen sich verschiedene empirische Zugänge zur Untersuchung dieser Forschungskompetenz identifizieren, die zwar inhaltliche Bezüge zu den Bildungswissenschaften aufweisen, jedoch vornehmlich durch an diesen Bereich angrenzende Disziplinen geprägt sind (Groß Ophoff et al., im Druck). So beschreiben Brown et al. (2010) für naturwissenschaftliche Kompetenz im schulischen Bereich evidenzbasiertes Schlussfolgern als Prozess, in dem auf eine bestimmte Prämisse bezogene Informationen zunächst analysiert, dann interpretiert und die resultierenden Folgerungen auf eine konkrete Problemstellung angewendet werden. Die am Anfang von Forschungsprozessen stehende Fähigkeit, Informationsbedarf zu erkennen, adäquate Forschungsfragen zu formulieren sowie gezielt und reflektiert zu recherchieren, wird unter dem Schlagwort *Informationskompetenz* untersucht (Catts & Lau, 2008) – beispielsweise indem für ein vorgegebenes Thema angemessene Hypothesen oder semantisch adäquate Schlagwörterkombinationen identifiziert werden müssen. In der Forschung zu *statistischer Kompetenz* wird die Fähigkeit thematisiert, statistische Informationen aus Tabellen, Grafiken oder Ergebnisdarstellungen systematisch analysieren und evaluieren zu können (Ben-Zvi & Garfield, 2004). Die Interpretation solcher Evidenz ermöglicht die Ableitung und Anwendung angemessener Schlussfolgerungen bzw. die kritische Reflexion von Ergebnisinterpretationen, wie sie typischerweise in der Diskussion von Forschungsartikeln anzutreffen sind. Diese Kompetenzfacette wird hauptsächlich als die Fähigkeit zum *kri-*

tischen Denken erforscht (Halpern, 1999). Die drei genannten Kompetenzfacetten Informationskompetenz, statistische Kompetenz und auf Forschung bezogenes kritisches Denken eignen sich nach Groß Ophoff et al. (im Druck) als strukturgebende Dimensionen von BFK.

Um die Qualität eines entsprechenden Testverfahrens beurteilen zu können, ist das Gütekriterium der Validität von entscheidender Bedeutung (Jenßen, Dunekacke & Blömeke, 2015, in diesem Beiheft). So ist über die Analyse der theoretisch postulierten Dimensionalität eines Konstrukts hinaus unter anderem zu prüfen, inwiefern dieses sich von angrenzenden Konstrukten erwartungskonform abgrenzen lässt (Konstruktvalidität; Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Newton & Shaw, 2014). Dazu werden Zusammenhänge mit Merkmalen analysiert, die konzeptuell entweder verwandt (konvergente Validität, z. B. selbsteingeschätzte Kompetenz) oder weiter entfernt sind (diskriminante Validität, z. B. fluide Intelligenz).

Wenngleich die Differenzierung von Kompetenz und Intelligenz nach Weinert (2001) allgemein als erforderlich erachtet wird, wird die konzeptuelle und empirische Trennbarkeit kontrovers diskutiert (Rindermann, 2006). Für die konzeptuelle Unterscheidung sind zwei Aspekte bedeutsam, zum einen die Spezifität und zum anderen die Erlern-/Förderbarkeit. Während sich mit Blick auf die Spezifität Kompetenzen auf klar umrissene Domänen beziehen (z. B. bildungswissenschaftliche Forschung), setzt fluide Intelligenz kein bereichsspezifisches Wissen voraus, sondern beschreibt die generalisierte Fähigkeit, durch Schlussfolgerungen neue kognitive Anforderungen spontan bewältigen zu können (Carpenter, Just & Shell, 1990). Sie wird typischerweise diagnostiziert, indem in figuralen (z. B. CFT 20 R; Weiß, 2006), numerischen und/oder verbalen (z. B. I-S-T 2000 R; Liepmann, Beauducel, Brocke & Amthauer, 2001) Denkaufgaben Regeln erkannt bzw. angewendet werden sollen. Betrachtet man den zweiten Unterscheidungsaspekt, zeigt sich, dass Kompetenz auf bereichsspezifischem Vorwissen basiert, das z. B. im Rahmen von (Hoch-)Schulbildung erworben und gezielt gefördert werden kann. Fluide Intelligenz gilt dagegen als stabiles Persönlichkeitsmerkmal (Neisser et al., 1996), das nicht veränderbar ist. Die Bedeutung der Unterscheidung von BFK und fluider Intelligenz für eine diskriminante Validierung liegt auf der Hand: Nach Brown et al. (2010) kann erst von Evidenz gesprochen werden, wenn Informationen systematisch analysiert und interpretiert werden können, wofür die Fähigkeit schlussfolgernden Denkens grundlegend ist. Zusätzlich sollten aber durch eine universitäre Ausbildung bereichsspezifisch Wissen und Strategien vermittelt werden, welche die Auseinandersetzung mit (bildungswissenschaftlicher) Evidenz unterstützen – dies ist Grundgedanke des sogenannten Value-Added Assessment (Liu, 2011).

Befunde aus Schulleistungsstudien sprechen ebenfalls dafür, Kompetenz und Intelligenz zu unterscheiden. So werden beispielsweise in PISA Kompetenzen als eindeutig abgrenzbar von allgemeinen kognitiven Fähigkeiten konzipiert und entsprechend auch getrennt erhoben (Baumert, Lüdtke, Trautwein & Brunner, 2009; Prenzel, Walter & Frey, 2007). Weiterhin finden sich in der Literatur Hinweise auf mindestens mittlere Zusammenhänge verschiedener Intelligenzfacetten mit Mathematikleistung (Taub, Floyd, Keith & McGrew, 2008) oder Leseleistung (Ramseier & Brühwiler, 2003) – in letzte-

rem Fall selbst unter Berücksichtigung weiterer Variablen wie Geschlecht, Schultyp oder sozial-kultureller Herkunft. Für den Hochschulbereich zeigt sich keine einheitliche Befundlage. So konnte Trapmann (2008) keine Zusammenhänge zwischen Studienerfolg und verbaler bzw. numerischer Verarbeitungskapazität nachweisen, was die Autorin damit erklärt, dass mit fortschreitender Bildung der Zusammenhang zurückzugehen scheint. Dagegen finden sich für figurale Subtests kognitiver Leistungstests kleine bzw. für verbale Subtests mittlere Zusammenhänge mit dem Studienerfolg (Giesen, Gold, Hummer & Jansen, 1986). Substanzielle Zusammenhänge werden dagegen berichtet zwischen fluider Intelligenz und der Fähigkeit zum komplexen Problemlösen, bei dem angesichts einer Vielzahl interagierender Informationen Komplexität reduziert werden muss (Süß, 2003).

Während die Analyse des Zusammenhangs zwischen Kompetenz und Intelligenz eine Form diskriminanter Validierung darstellt, ermöglicht die Analyse des Zusammenhangs zwischen unterschiedlichen Erhebungsmethoden des gleichen Konstrukts, also z. B. objektiver Testleistung und Selbsteinschätzung, die Prüfung konvergenter Validität. Die Selbsteinschätzung eigener Fähigkeiten ist definiert als die Wahrnehmung der Person ihrer aktuellen Kompetenz bezüglich einer bestimmten Tätigkeit (Wigfield & Eccles, 2000). Im Sinne des Multitrait-Multimethod-Ansatzes (Campbell & Fiske, 1959) wären – im Gegensatz zur Abgrenzung von fluider Intelligenz – hohe Zusammenhänge zu objektiv gemessener Kompetenz zu erwarten. Allerdings ist speziell die Befundlage in diesem Bereich kontrovers: So wurden in schulischen Stichproben allenfalls kleine bis mittlere Zusammenhänge zwischen den Erhebungsmethoden gefunden (Freiberger, Steinmayr & Spinath, 2012). Auch Studierende scheinen sich bzgl. ihrer Kompetenz zu verschätzen (Chevalier, Gibbons, Thorpe, Snell & Hoskins, 2009).

Basierend auf den dargestellten Befunden stellt dieser Beitrag Analysen zur Konstruktvalidität eines Testinstruments zur Erfassung von BFK vor, das im Verbundprojekt Learning the Science of Education (Schladitz et al., 2013) entwickelt wurde. Es wird erwartet, dass die BFK

a) im Sinne diskriminanter Validierung kleine bis mittlere Zusammenhänge zur fluiden Intelligenz sowie
b) im Sinne konvergenter Validierung mittlere bis hohe Zusammenhänge zur selbsteingeschätzten Kompetenz aufweist.

Entsprechend der inhaltlichen Ausrichtung der Kompetenzfacetten sowie der Konzeption der Intelligenzmaße und der Skala zur Selbsteinschätzung werden unterschiedlich starke Effekte erwartet. Die Zusammenhänge zwischen der sprachfreien Messung der Intelligenz und den Kompetenzfacetten sollten in allen Fällen im mittleren Bereich liegen. Für die verbale Messung der Intelligenz wird erwartet, dass mittlere Zusammenhänge zu denjenigen Kompetenzfacetten bestehen, die einen hohen Anteil sprachlicher Informationen beinhalten (Informationskompetenz, forschungsbezogenes kritisches Denken), und kleine Zusammenhänge zu statistischer Kompetenz.

Die neu entwickelte Skala zur selbsteingeschätzten BFK sollte dieselben Facetten abbilden wie der Kompetenztest. Entsprechend wird erwartet, dass zwischen den jeweils gleichen Facetten höhere Zusammenhänge bestehen als zwischen den nicht zusammengehörigen Facetten. Im Sinne diskriminanter Validierung werden keine Zusammenhänge zwischen selbsteingeschätzter BFK und fluider Intelligenz erwartet.

2. Methoden

2.1 Instrumente

Der folgende Abschnitt stellt die für diese Studie relevanten Instrumente vor. Zusätzlich wurden noch weitere Skalen eingesetzt (z. B. Motivationsskalen), die für den vorliegenden Beitrag nicht relevant sind.

Kompetenztest

Das Instrument zur Erfassung der BFK wurde im Rahmen des Projekts neu entwickelt. Anhand bestehender Literatur im Bereich Forschungskompetenz wurden die Inhaltsbereiche Informationskompetenz, statistische Kompetenz sowie forschungsbezogenes kritisches Denken als relevante Facetten identifiziert und als Konstruktionsheuristik für die Itementwicklung genutzt (Groß Ophoff et al., im Druck). Zur Gewährleistung der Inhaltsvalidität wurden die ersten Aufgabenentwürfe mit ExpertInnen ($N = 5$) auf dem Gebiet der Kompetenzmessung bzw. Pädagogischen Psychologie diskutiert. Nach einer Voruntersuchung ($N = 6$ Studierende) bezüglich der Verständlichkeit der Items bestand der Itempool aus 226 Einzelitems (Beispiele s. Abb. 1).

Die postulierte Multidimensionalität und die angestrebten Skalenreliabilitäten bedingten eine sehr große Itemanzahl, weshalb für die Datenerhebung ein unvollständiges Blockdesign (Frey, Hartig & Rupp, 2009) verwendet wurde. Die Items wurden auf Blöcke zu je acht Aufgaben aufgeteilt. Diese wurden so zusammengestellt, dass die Bearbeitungszeit etwa gleich war, alle Facetten auf Aufgabenebene gleichmäßig vertreten und keine Aufgaben mit dem gleichen Itemstamm in einem Block enthalten waren, um Redundanzen und lokale Abhängigkeiten zu vermeiden. Jeweils vier Aufgabenblöcke wurden zu einem Testheft zusammengefasst (Abb. 2), sodass jeder Person 32 Testaufgaben vorlagen. Um Reihenfolgeeffekte einzuschränken, wurde jedes Testheft zur Hälfte in blockinverser Reihenfolge dargeboten.

Es wurden 23 Items zu Informationskompetenz, 68 zu statistischer Kompetenz und 91 zu forschungsbezogenem kritischen Denken in die Analysen einbezogen. Die Diskrepanz zur ursprünglichen Itemanzahl liegt darin begründet, dass nur Items im Multiple-Choice- und Fill-in-Format in die aktuelle Analyse aufgenommen wurden sowie Items mit nicht adäquater Modellpassung im Vorfeld ausgeschlossen wurden (Groß Ophoff et al., im Druck).

Um bei einer Datenbankrecherche Suchwörter miteinander zu verbinden, können die logischen Operatoren AND, OR bzw. NOT verwendet werden, die zu unterschiedlichen Suchergebnissen führen.

Bringen Sie die logischen Kombinationen mit den Zahlen 1 (= meiste Suchergebnisse) bis 3 (wenigste Suchergebnisse) in die richtige Reihenfolge.

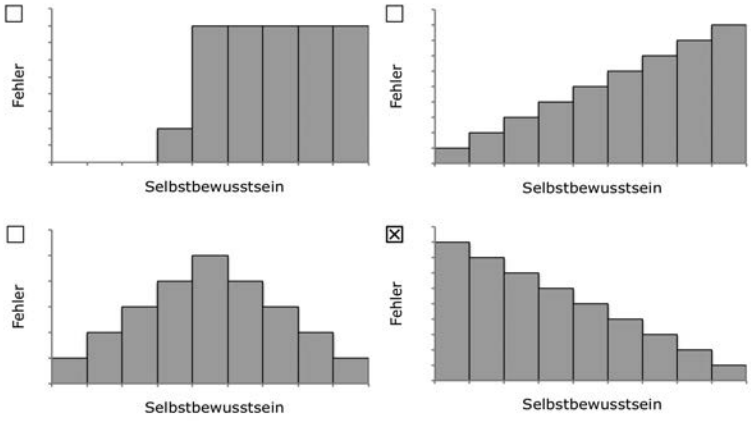
a) ___ Heterogenität

b) ___ Heterogenität OR Grundschule

c) ___ Heterogenität NOT Grundschule

In Experimenten hat sich gezeigt, dass ein höheres Selbstbewusstsein die Fehleranzahl in Diktaten verringert.

Welche der folgenden Verteilungsformen bildet dieses Ergebnis ab?



Sie lesen die folgenden beiden Kurzdarstellungen von Forschungsbefunden:

Befund A: Die Notwendigkeit einer grundlegenden Reform des Bildungssystems wird durch die nachgewiesene hohe Unzufriedenheit der untersuchten Berufsgruppen (Lehrer/innen, Pädagog/inn/en, Erzieher/innen) belegt. Zukünftig müssen die Struktur und die Prozesse in Bildungseinrichtungen stärker auf die Bedürfnisse der beteiligten Berufsgruppen abgestimmt werden, da eine Verbesserung der Ausbildungsqualität unbedingt erforderlich ist.

Befund B: Der im Gruppenvergleich nachgewiesene positive Effekt der Bildung von Lehrerteams auf die Unterrichtsqualität unterstützt die Annahme, dass durch Ausbildungsteams positive Effekte erzielt werden können. Zukünftig sollte die Entwicklung und Prüfung der Effekte konkreter Teambildungsmaßnahmen in Schulen weiter untersucht werden.

Geben Sie an, welche Merkmale eher für Befund A bzw. Befund B zutreffen:

	Eher für A zutreffend	Für A und B in ähnlichem Maß zutreffend	Eher für B zutreffend
Die Ableitungen der Studie sind direkt auf die Untersuchungsergebnisse bezogen.	☐	☐	☒
Die untersuchten Merkmale werden konkret benannt.	☐	☒	☐
Die verwendete Untersuchungsanlage wird deutlich.	☐	☐	☒

Abb. 1: Beispielitems zu den Kompetenzfacetten Informationskompetenz (oben), statistische Kompetenz (Mitte) und forschungsbezogenes kritisches Denken (unten)

Testheft	1		2		3		…	18		19		20	
	V	R	V	R	V	R		V	R	V	R	V	R
	1	4	2	5	3	6		18	1	19	2	20	3
	2	3	3	4	4	5		19	20	20	1	1	2
	3	2	4	3	5	4		20	19	1	20	2	1
	4	1	5	2	6	3		1	18	2	19	3	20

Abb. 2: Verteilung der Aufgabenblöcke 1–20 á 8 Aufgaben im Testheftdesign mit 20 Testheften, 4 Blockpositionen pro Testheft; V = Vorwärtsreihenfolge der Blöcke, R = Rückwärtsreihenfolge

Intelligenztests

Zur Erfassung der fluiden Intelligenz wurden Subskalen aus zwei verschiedenen Tests eingesetzt. Die Subskala „Gemeinsamkeiten finden" aus dem Intelligenz-Struktur-Test (I-S-T 2000 R; Liepmann et al., 2001) erfasst verbale Intelligenz. In den für die Studie relevanten Alterskohorten werden im Skalenhandbuch folgende Mittelwerte und Standardabweichungen für die aus 20 Items bestehende Subskala angegeben: 15–20 Jahre: 10.72 (SD = 3.49); 21–25 Jahre: 11.12 (SD = 3.67); 26–30 Jahre: 10.39 (SD = 3.92); 31–40 Jahre: 9.83 (SD = 3.98). Zusätzlich wurde die Subskala „Reihenfortsetzen" aus dem Culture Fair Intelligence Test (CFT 20 R; Weiß, 2006) eingesetzt, um neben der sprachabhängigen (= verbalen) auch eine sprachfreie Messung fluider Intelligenz zu erhalten. Je einem Drittel der Personen wurde der CFT 20 R oder der I-S-T 2000 R zur Bearbeitung vorgelegt, sodass jede/r TeilnehmerIn nur einen der Tests bearbeitet hat. Der CFT 20 R bietet neben den Original-Subskalen jeweils auch eine Parallelversion (Testform A und B). Um die Testzeit für alle Personen gleichzuhalten, wurden hier beide Parallelversionen der Subskala verwendet.

Skala zur selbsteingeschätzten BFK

Es wurde eine Skala mit acht Items neu entwickelt, die sich auf Aspekte bezieht, die auch im Kompetenztest gemessen werden: Literaturrecherche/Fragenstellen (vgl. Informationskompetenz; Items s. Tab. 3), Umgang mit Methoden-/Ergebnisdarstellungen (vgl. statistische Kompetenz) sowie kritische Reflexion von Befunden (vgl. forschungsbezogenes kritisches Denken). Zusätzlich wurde ein Globalitem zum Verständnis bildungswissenschaftlicher Studien gestellt. Die Einschätzung erfolgte auf einer fünfstufigen Likert-Skala.

2.2 Stichprobe

Es konnten 1360 Personen an sechs deutschen Hochschulen für die Teilnahme gewonnen werden. Die Merkmale der (Teil-)Stichproben sind in Tabelle 1 aufgeführt.

	Gesamtstichprobe	Teilstichprobe „Gemeinsamkeiten finden“ (verbale Intelligenz)	Teilstichprobe „Reihenfortsetzen“ (sprachfreie Intelligenz)
N	1360	461	434
Alter *(M, SD)*	22.93 (3.95)	22.74 (3.70)	22.88 (3.30)
Semester (Median)	3	3	3
weiblich (%)	75.4	75.6	75.5
Studiengänge (%)			
Lehramt	61.8	63.7	61.8
Erziehungswissenschaft	23.1	22.7	22.7
Gesundheitspädagogik	5.1	4.5	5.5
andere	8.9	7.7	9.0

Anmerkungen. N = Stichprobengröße; *M* = Mittelwert; *SD* = Standardabweichung.

Tab. 1: Deskriptive Statistiken der Gesamt- und Teilstichproben

2.3 Analyseverfahren

Basis für die Zusammenhangsanalysen sind Skalen, die den gängigen psychometrischen Gütekriterien genügen (Adams, 2002; Schermelleh-Engel, Moosbrugger & Müller, 2003). In diesem Sinne wurden zunächst die BFK (Groß Ophoff, Schladitz & Wirtz, 2014) sowie die Intelligenzmaße und die selbsteingeschätzte Kompetenz skalenanalytisch evaluiert. Für jede der Kompetenzfacetten (Informationskompetenz, statistische Kompetenz, forschungsbezogenes kritisches Denken) wurden fünf Personenschätzer (Plausible Values; von Davier, Gonzalez & Mislevy, 2009) ermittelt, die als Kriterium für die vorgestellten strukturanalytischen Analysen dienten. Im Rahmen der vorliegenden Analyse wurde für die dichotomen Daten aus den Intelligenzskalen auf Messmodellebene die Passung zu den Annahmen des Raschmodells (1PL-Modell) und des Birnbaum-Modells (2PL-Modell) auf Basis der Informationskriterien Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) sowie Consistent Akaike Information Criterion (CAIC) kontrastierend geprüft. Niedrigere Werte zeigen hierbei eine bessere Datenkompatibilität unter Berücksichtigung der Modellkomplexität an (Schermelleh-Engel et al., 2003). Bei der anschließenden Überprüfung der Itemeigenschaften wurden Items beibehalten, die anhand des Weighted Mean Square (WMNSQ) und des dazugehörigen t-Werts gut zur Gesamtskala passten. Die Werte sollten im Bereich $0.80 \leq \text{WMNSQ} \leq 1.20$ liegen und keine Signifikanz aufweisen. Diese Werte wurden von Linacre (1994) als stichprobengrößenfaire Grenzwerte vorgeschlagen, um größere Stichprobenumfänge nicht zu bestrafen. Die EAP/PV (expected a posteriori/plausible value)-Reliabilität, die mit Cronbach's Alpha vergleichbar ist, gibt die Messgenauigkeit der Skala an.

Zur Analyse der kontinuierlichen Daten aus der Skala zur selbsteingeschätzten Kompetenz wurde zunächst mit einem explorativen Strukturgleichungsmodell (Asparouhov & Muthén, 2009) die Faktorenstruktur analysiert und anschließend die am besten zu den Daten passende Struktur konfirmatorisch modelliert. Die Modellgüte wurde anhand der gängigen Maße überprüft, wonach der Root Mean Square Error of Approximation (RMSEA) höchstens .08 und der Comparative-Fit-Index (CFI) mindestens .95 betragen sollte (Schermelleh-Engel et al., 2003).

Nach Sicherstellung aller Skalengütekriterien wurden die Zusammenhänge zwischen Kompetenzfacetten, Intelligenzkonstrukten sowie selbsteingeschätzter Kompetenz analysiert und ebenfalls anhand der o. g. Gütekriterien bewertet.

3. Ergebnisse

3.1 Psychometrische Prüfung der Intelligenzskalen

Der Modellvergleich für die Intelligenzskalen ergab für beide Konstrukte eine bessere Passung des 2PL-Modells gegenüber dem Rasch-Modell (Tab. 2). Aus diesem Grund wurde für alle weiteren Analysen das 2PL-Modell verwendet.

Betrachtet man die psychometrischen Eigenschaften der einzelnen Items, zeigen sich für die Skala „Gemeinsamkeiten finden“ aus dem Test zur verbalen Intelligenz Lösungshäufigkeiten zwischen 4.99 % und 93.26 % mit einer durchschnittlichen Lösungshäufigkeit von 57.60 %. Die Skala ist so konzipiert, dass die Aufgaben im Verlauf der Skala schwieriger werden. Dies bildet sich auch im vorliegenden Fall ab: In der ersten Hälfte der Skala liegt die durchschnittliche Lösungshäufigkeit bei 78.64 %, in der zweiten Hälfte bei 36.55 %. Die Werte des WMNSQ liegen mit [0.99; 1.05] durchgehend im guten Bereich und weisen keine Signifikanz auf. Die aus dem 2PL-Modell ermittelten Schwierigkeiten der Items liegen im Bereich [−7.12; 23.85] mit Diskriminationsparametern von [0.12; 3.25]. Die EAP/PV-Reliabilität für diese Skala kann mit .73 als zufriedenstellend angesehen werden.

	„Gemeinsamkeiten finden“ (verbale Intelligenz)		**„Reihenfortsetzen“ (sprachfreie Intelligenz)**	
	1 PL	**2 PL**	**1 PL**	**2 PL**
AIC	9058.651	**8887.220**	9161.037	**9046.856**
BIC	9145.452	**8056.690**	9275.082	**9270.873**
CAIC	9166.452	**9097.690**	9303.082	**9325.873**

Anmerkungen. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion. Das jeweils am besten passende Modell wurde hervorgehoben.

Tab. 2: Modellvergleiche für die Intelligenzskalen

Der Mittelwert korrekt gelöster Items unterschied sich mit 11.52 (*SD* = 2.97) von den Normwerten der Skala „Gemeinsamkeiten finden" aus dem I-S-T 2000 R. So lagen Studierende in der Altersgruppe von 15 bis 20 Jahren (*n* = 104) mit einem Mittelwert von 11.93 (95 %-Konfidenzintervall: 11.44; 12.42, vgl. 3.1) signifikant über dem der Referenzgruppe und streuten entsprechend der Standardabweichung von 2.56 (95 %-Konfidenzintervall: 2.21; 2.91) signifikant geringer. Die größte Altersgruppe der 21- bis 25-Jährigen (*n* = 301) unterschied sich mit einem Mittelwert von 11.51 nicht von der Referenzgruppe, erwies sich aber mit einer Standardabweichung von 2.97 (95 %-Konfidenzintervall: 2.74; 3.21) ebenfalls als signifikant homogener.

Aus der Skala „Reihenfortsetzen" sind die WMNSQ-Werte mit [0.89; 1.09] allesamt zufriedenstellend und nicht signifikant. Die durchschnittliche Lösungshäufigkeit liegt bei 78.24 % mit Werten im Bereich [32.03 %; 97.23 %]. Auch in diesem Fall zeigt sich die in der Testentwicklung intendierte Steigerung der Aufgabenschwierigkeit über die Skala hinweg. Halbiert man die beiden Testformen A und B, ergeben sich für die ersten Hälften durchschnittliche Lösungshäufigkeiten von 89.95 % bzw. 91.67 % und für die zweiten Hälften von 76.27 % bzw. 51.50 %. Die ermittelten Schwierigkeiten liegen damit korrespondierend im Bereich [−2.60; 0.83], die Diskriminationsparameter zwischen [0.56; 5.35]. Insgesamt weist die Skala mit EAP/PV = .71 eine zufriedenstellende Reliabilität auf.

3.2 Struktur der Itemgruppen zur „selbsteingeschätzten Kompetenz"

Die Analyse des explorativen Strukturgleichungsmodells lieferte Hinweise auf eine zweidimensionale Struktur (Tab. 3). Den ersten Faktor bilden die vier Items aus den Bereichen Recherche/Fragenstellen und kritische Reflexion, die drei Items des zweiten Faktors beziehen sich auf das Lesen von Methoden-/Ergebnisdarstellungen. Das Item zum globalen Verständnis wurde ausgeschlossen. Die beiden Dimensionen weisen jeweils zufriedenstellende Reliabilität auf ($\alpha = .76$ bzw. $\alpha = .70$) und korrelieren latent mit .70 miteinander. Die Gütekriterien für das konfirmatorische Strukturmodell liegen ebenfalls im akzeptablen Bereich (RMSEA = .066; CFI = .970).

3.3 Zusammenhang zwischen Kompetenzfacetten, Intelligenzkonstrukten und selbsteingeschätzter Kompetenz

In Strukturgleichungsmodellen wurden anschließend die Zusammenhänge zwischen Kompetenzfacetten, Intelligenzkonstrukten und der selbsteingeschätzten Kompetenz analysiert (s. Abb. 3, aus Gründen der Übersichtlichkeit sind die Messmodelle nicht abgebildet). Um die lokale Abhängigkeit der Schätzungen aus denselben Simulationsläufen zu berücksichtigen, wurden simulationsspezifische Residualkorrelationen der manifesten Plausible Values definiert. Als Schätzalgorithmus wurde Mean and Variance Adjusted Weighted Least Squares (WLSMV; z. B. Beauducel & Herzberg, 2006) ver-

Item	λ	SE	R^2
Dimension 1 – Literaturrecherche/Fragenstellen & kritische Reflexion			
Ich fühle mich sicher bei der Literaturrecherche für wissenschaftliche Arbeiten.	.47	.05	.24
Ich fühle mich sicher in der Formulierung wissenschaftlicher Fragestellungen und Hypothesen.	.74	.05	.52
Ich kann Forschungsergebnisse kritisch reflektieren.	.66	.05	.47
Ich kann die Qualität einer bildungswissenschaftlichen Studie sicher beurteilen.	.59	.03	.34
Dimension 2 – Lesen von Methoden-/Ergebnisdarstellungen			
Ich kann Daten aus Diagrammen, Tabellen und Texten gut erfassen und sicher interpretieren.	.70	.05	.46
Ich kann methodische und statistische Aspekte aus bildungswissenschaftlichen Studien gut verstehen.	.80	.03	.64
Ich kann Ergebnisse einer bildungswissenschaftlichen Studie angemessen interpretieren.	.55	.04	.50

Anmerkungen. λ = standardisierte Faktorladung; SE = Standardfehler; R^2 = Varianzaufklärung. Für alle Ladungen gilt $p < .001$.

Tab. 3: Faktorladungen, Fehlerterme und Varianzaufklärung der Items zur selbsteingeschätzten BFK

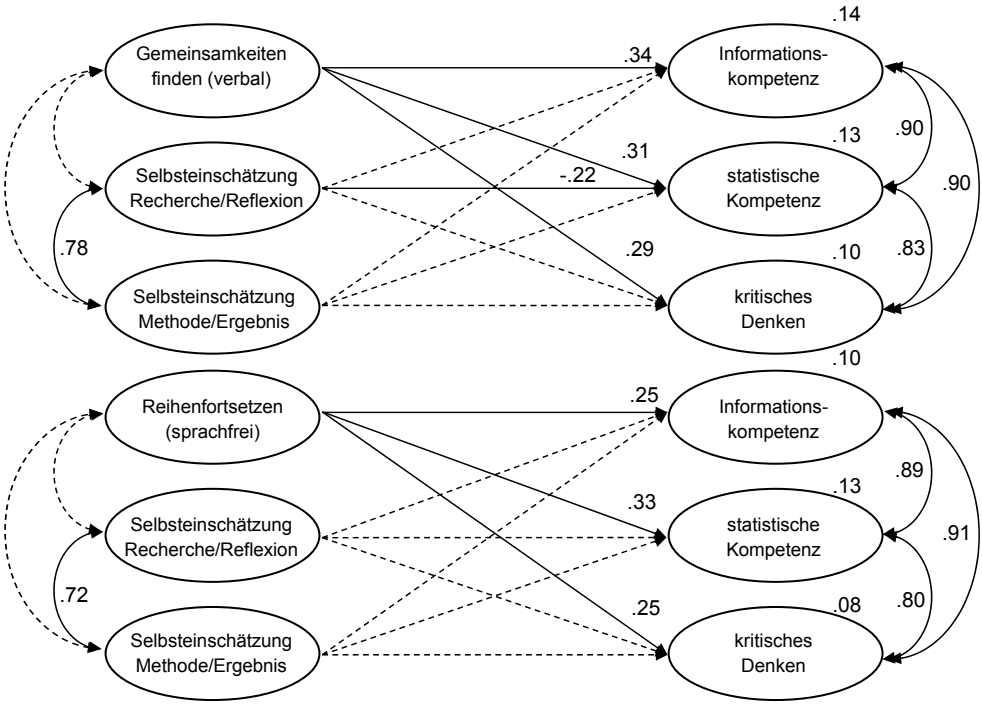


Abb. 3: Strukturmodelle zum Zusammenhang zwischen den Intelligenzskalen, der selbsteingeschätzten Kompetenz und den Kompetenzfacetten. Dargestellt sind signifikante standardisierte Regressionsgewichte, latente Korrelationen sowie die Varianzaufklärung. Nichtsignifikante Pfade sind gestrichelt dargestellt.

wendet, wobei fehlende Werte listenweise ausgeschlossen wurden. Die Maße der Modellgüte waren in beiden Fällen gut bis zufriedenstellend (Skala „Gemeinsamkeiten finden“: RMSEA = .020; CFI = .966; Skala „Reihenfortsetzen“: RMSEA = .019; CFI = .962). Die dargestellten vollständig standardisierten Regressionsgewichte lagen für die Intelligenzkonstrukte im schwachen bis mittleren Bereich (Cohen, 1988). Die Effekte der selbsteingeschätzten Kompetenz erwiesen sich hingegen bis auf einen Fall als nicht bedeutsam ($|\beta| \leq .18$).

4. Diskussion

Da dieser Beitrag mit dem Ziel konvergenter und diskriminanter Validierung der BFK verfasst wurde, lag der Fokus auf den Zusammenhängen mit fluider Intelligenz sowie der Selbsteinschätzung der Forschungskompetenz. Die Darstellung der detaillierten Skalenanalyse und Struktur der BFK ist für einen weiteren Beitrag vorgesehen (erste Ergebnisse bei Groß Ophoff et al., im Druck), ebenso wie Befunde zu Korrelationen mit weiteren, im Rahmen des Forschungsprojekts LeScEd erhobenen Konstrukten.

Die vorangehende Überprüfung der Struktur der Intelligenzskalen zeigt, dass die Analyse mittels IRT-Methoden eine wertvolle Ergänzung zur ursprünglichen Konstruktion anhand klassischer Testtheorie darstellt. So konnte an vergleichsweise großen Stichproben verdeutlicht werden, dass die ermittelten Reliabilitäten grenzwertig zufriedenstellend sind und durch weitere Itemanalysen ggf. noch verbessert werden können. Das besser zu den Daten passende 2PL-Modell hat gegenüber dem Rasch-Modell den Nachteil, dass die identifizierten Dimensionen schwieriger zu interpretieren sind (Rauch & Hartig, 2012). Einschränkend ist außerdem zu berücksichtigen, dass die hier untersuchte studentische Stichprobe z. T. signifikant höhere Mittelwerte und eine geringere Varianz in der Skala „Gemeinsamkeiten finden“ der verbalen Intelligenz aufweist als die im Manual wiedergegebenen Normierungsstichproben (vgl. 3.1; Liepmann et al., 2001), was zu einer Verzerrung der Effektschätzungen führen kann (Urban & Mayerl, 2008). Eine erneute Überprüfung der Intelligenzskalen anhand probabilistischer Testmodelle und der Passung des Rasch-Modells in weniger spezifischen Stichproben sollte daher in Erwägung gezogen werden.

Die entwickelte Skala zur selbsteingeschätzten BFK zeigt vielversprechende Gütekriterien. Man muss allerdings berücksichtigen, dass die Struktur nicht deckungsgleich mit den Facetten des Kompetenztests ist. In der Analyse der Skalenstruktur bildeten sich zwei statt drei Inhaltsbereiche ab. Es scheint, als gäbe es Unterschiede hinsichtlich der Verarbeitung hauptsächlich verbaler (Informationskompetenz, forschungsbezogenes kritisches Denken) und hauptsächlich statistischer Informationen (statistische Kompetenz). Die Nähe zu Kompetenzkonstrukten aus der Schulleistungsforschung (Lesekompetenz, mathematische Kompetenz) liegt auf der Hand und sollte in künftigen Studien explizit berücksichtigt werden.

4.1 Diskriminante und konvergente Validierung

Die Zusammenhänge zwischen der BFK und der Intelligenz (diskriminante Validierung) liegen wie erwartet im mittleren Bereich. Die somit schwächeren Zusammenhänge, als sie z. B. in PISA ermittelt wurden (Baumert et al., 2009; Prenzel et al., 2007), gehen einher mit Trapmanns (2008) Standpunkt, dass der Zusammenhang zwischen Intelligenz und akademischer Leistung im Verlauf des Bildungsweges abnimmt. Es zeigt sich also, dass BFK und fluide Intelligenz (mit den Subskalen „Gemeinsamkeiten finden" und „Reihenfortsetzen") zwar verwandte, aber empirisch gut voneinander abgrenzbare Konstrukte darstellen. Die auf Vorwissen und Lernerfahrungen basierende Forschungskompetenz stellt offensichtlich ein über die fluide Intelligenz hinausgehendes Merkmal dar, was für eine getrennte Erfassung beider Bereiche spricht.

Aufseiten der selbsteingeschätzten BFK (konvergente Validierung) zeigen sich dagegen unerwartet fast keine Zusammenhänge zur objektiv gemessenen BFK. Die einzige signifikante Korrelation erweist sich sogar als negativ, wonach eine selbsteingeschätzte höhere Kompetenz in Recherche und Reflexion von Evidenz mit einer geringeren statistischen Kompetenz einhergeht. Dies kann ein Hinweis darauf sein, dass mit einer allgemeinen, also nicht auf ein spezifisches Lehrangebot bezogenen Selbsteinschätzung von Kompetenz eher Präferenzen für bestimmte Aspekte der Beschäftigung mit bildungswissenschaftlicher Forschung gemessen werden – wobei der Effekt eher klein ausfällt und nicht überinterpretiert werden sollte. Die nicht aufzeigbaren Zusammenhänge zwischen objektiver und subjektiver BFK können darin begründet sein, dass die selbsteingeschätzte Kompetenz tatsächlich keinen geeigneten Prädiktor für die objektive Kompetenzausprägung darstellt. Damit bestätigt sich die bereits angedeutete Diskrepanz zwischen Validierungstheorie (Campbell & Fiske, 1959) und empirischer Befundlage (z. B. Chevalier et al., 2009): Trotz Messung desselben Konstrukts – was für einen hohen Zusammenhang spräche – scheinen die beiden Operationalisierungen voneinander unabhängig. Eine generelle Schwierigkeit beim Einsatz von Selbsteinschätzungsskalen ist deren Losgelöstheit von tatsächlichen Kompetenzaufgaben. So ist denkbar, dass sich bei Vorgabe der spezifischen Kompetenzitems mit der Frage nach der selbst eingeschätzten Lösungswahrscheinlichkeit andere Ergebnisse zeigen würden. Die abstrakte Bewertung recht weit formulierter Kompetenzen bietet evtl. ein hohes Potential für eine falsche Einschätzung des eigenen Kompetenzniveaus. Es empfiehlt sich daher, für eine zuverlässige Kompetenzerfassung nicht ausschließlich Instrumente zur Selbsteinschätzung (z. B. Braun, Gusy, Leidner & Hannover, 2008) einzusetzen, sondern, wenn möglich, eine objektive Kompetenzmessung zu ermöglichen.

Trotz der nachweislich besser passenden dreidimensionalen Kompetenzstruktur (Groß Ophoff et al., 2014) zeigen sich kaum differenzierte Zusammenhänge zwischen Intelligenz und Selbsteinschätzung auf der einen sowie den drei Kompetenzfacetten auf der anderen Seite. In Verbindung mit den sehr hohen Interkorrelationen der Facetten wirft dieser Umstand die Frage auf, inwiefern es praktisch sinnvoll (z. B. für die Rückmeldung an Studierende im Rahmen von Lehrveranstaltungen) ist, BFK so aus-

führlich zu differenzieren. Auch wenn z. T. in Large-Scale-Studien trotz vergleichbarer oder sogar höherer Zusammenhänge eine Trennung von Teilkompetenzen als Beitrag zur grundlegenden Erforschung des Konstrukts beibehalten wird (Artelt & Schlagmüller, 2004), sollten vertiefende Analysen zur weiteren Klärung beitragen. Letztlich bleibt es vermutlich eine inhaltliche, mit Blick auf den Verwendungszusammenhang (i. S. von Consequential Validity; vgl. Newton & Shaw, 2014) zu treffende Entscheidung für oder gegen die Unterscheidung der Kompetenzfacetten.

Abschließend lässt sich festhalten, dass die diskriminante Validierung der BFK anhand der fluiden Intelligenz vorgenommen werden konnte. Mit der selbsteingeschätzten Kompetenz im Sinne konvergenter Validierung dagegen wird offensichtlich nicht das gleiche Konstrukt, sondern eventuell individuelle Präferenzen bei der Auseinandersetzung mit bildungswissenschaftlicher Forschung gemessen.

4.2 Ausblick auf zukünftige Forschung

Der Einbezug zweier Skalen aus verschiedenen Intelligenzbereichen (verbal und sprachfrei) trägt nicht zur weiteren Validierung bzw. Aufklärung des Konstrukts BFK bei. Jedoch lässt die aus testökonomischen Gründen vorgenommene Aufteilung der Intelligenzskalen auf zwei nichtüberlappende Teilstichproben keine Analyse inkrementeller Effekte der spezifischen Intelligenzarten zu. Eine gemeinsame Erfassung in einer Stichprobe sowie der Einbezug anderer Intelligenzkomponenten (z. B. numerische Intelligenz) können weitere Hinweise auf differenzierte Zusammenhangsstrukturen zur BFK liefern. Weiterhin werden differenzielle Itemeigenschaften (Zumbo, 2007) in Abhängigkeit von der Intelligenz untersucht, um genauere Informationen darüber zu erhalten, welche Items bzw. Itemgruppen besonders stark von der Intelligenzausprägung beeinflusst werden.

Neben der Intelligenz und der Selbsteinschätzung können sich weitere Personenmerkmale wie z. B. Motivation (z. B. Spinath & Steinmayr, 2012) als relevant für die Ausbildung der BFK erweisen. Weinert (2001) empfiehlt auch für solche nicht-kognitiven Merkmale eine getrennte Erfassung, um Zusammenhänge zur jeweiligen Kompetenz analysieren zu können. Darüber hinaus spielen vor allem in einem – im Gegensatz zur schulischen Ausbildung – relativ frei gestaltbaren Hochschulstudium Lerngelegenheiten (Blömeke, Suhl, Kaiser, Felbrich & Schmotz, 2010) eine Rolle. Diese Aspekte werden derzeit im Rahmen einer längsschnittlichen Untersuchung der BFK analysiert.

Die Befunde zur Abgrenzbarkeit der BFK von fluider Intelligenz bieten eine Grundlage für die Entwicklung gezielter Lehr- bzw. Weiterbildungsangebote. So können je nach Zielsetzung der Lehrveranstaltung ausgewählte Teilkompetenzen gefördert werden, die zwar mit allgemeiner Intelligenz zusammenhängen, aber auch darüber hinausgehend zu einer verbesserten Leistung im Umgang mit bildungswissenschaftlicher Forschung führen. Jedoch bleibt das Desiderat bestehen, inwiefern die Vermittlung entsprechender Kompetenzfacetten tatsächlich evidenzbasiertes Arbeiten und kontinuierliche Professionalisierung in der späteren beruflichen Praxis unterstützen kann. Hier

bieten sich Studien an, die den Übergang in den Beruf verfolgen (z. B. COACTIV-R; Kunter et al., 2011), die Kompetenzen Berufstätiger systematisch erfassen und idealerweise auch langfristig begleiten.

Literatur

Adams, R. J. (2002). Scaling PISA cognitive data. In R. J. Adams & M. L. Wu (Hrsg.), *PISA 2000 technical report* (S. 99–108). Paris: OECD.

Artelt, C., & Schlagmüller, M. (2004). Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 169–196). Wiesbaden: VS Verlag für Sozialwissenschaften.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397–438.

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4*(3), 165–176 [DOI: 10.1016/j.edurev.2009.04.002].

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of Maximum Likelihood versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(2), 186–203 [DOI: 10.1207/s15328007sem 1302_2].

Ben-Zvi, D., & Garfield, B. (2004). *The challenge of developing statistical literacy, reasoning and thinking*. New York: Kluwer Academic Publishers.

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

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

Braun, E., Gusy, B., Leidner, B., & Hannover, B. (2008). Kompetenzorientierte Lehrevaluation – Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp). *Diagnostica, 54*(1), 30–42.

Brown, N. J. S., Furtak, E. M., Timms, M., Nagashima, S. O., & Wilson, M. (2010). The Evidence-Based Reasoning Framework: Assessing scientific reasoning. *Educational Assessment, 15*(3/4), 123–141.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56*(2), 81–105.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404–431.

Catts, R., & Lau, J. (2008). *Towards information literacy indicators: Conceptual framework paper*. Paris.

Chevalier, A., Gibbons, S., Thorpe, A., Snell, M., & Hoskins, S. (2009). Students' academic self-perception. *Economics of Education Review, 28*(6), 716–727.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), S. 281–302.

Freiberger, V., Steinmayr, R., & Spinath, B. (2012). Competence beliefs and perceived ability evaluations: How do they contribute to intrinsic motivation and achievement? *Learning and Individual Differences, 22*, 518–522 [DOI: 10.1016/j.lindif.2012.02.004].

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53.

Giesen, H., Gold, A., Hummer, A., & Jansen, R. (1986). *Prognose des Studienerfolgs*. Frankfurt a. M.: Institut für Pädagogische Psychologie.

Groß Ophoff, J., Schladitz, S., Lohrmann, K., & Wirtz, M. (im Druck). Evidenzorientierung in bildungswissenschaftlichen Studiengängen: Entwicklung eines Strukturmodells zur Forschungskompetenz. In W. Bos, K. Drossel & R. Strietholt (Hrsg.), *Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen*. Münster: Waxmann.

Groß Ophoff, J., Schladitz, S., & Wirtz, M. (2014, März). *Struktur von Research Literacy in den Bildungswissenschaften*. Vortrag auf der Tagung der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE). Berlin.

Halpern, D. F. (1999). Teaching for critical thinking: Helping college students develop the skills and dispositions of a critical thinker. *New Directions for Teaching and Learning, 1999*, 69–74 [DOI: 10.1002/tl.8005].

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik, 61. Beiheft*, 11–31.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W., & Stanat, P. (Hrsg.) (2010). *PISA 2009. Bilanz nach einem Jahrzehnt*. Münster/New York/München/Berlin: Waxmann.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Überarbeitete Fassung des Antrags an die DFG auf Einrichtung eines Schwerpunktprogramms. *Zeitschrift für Pädagogik, 52*(6), 876–903.

KMK Kultusministerkonferenz (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16. 12. 2004*. http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf [29. 05. 2014].

KMK Kultusministerkonferenz (2005). *Qualifikationsrahmen für Deutsche Hochschulabschlüsse*. http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2005/2005_04_21-Qualifikationsrahmen-HS-Abschluesse.pdf [29. 05. 2014].

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster/New York/München/Berlin: Waxmann.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2001). *Intelligenz-Struktur-Test 2000R*. Göttingen: Hogrefe.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch measurement transactions, 7*(4), 328.

Liu, O. L. (2011). Value-added assessment in higher education: a comparison of two methods. *Higher Education, 61*(4), 445–461.

McMillan, J. H., & Schumacher, S. (2010). *Research in education. Evidence-based inquiry* (7. Aufl.). Upper Saddle River: Pearson.

Neisser, U., Boodoo, G., Bouchard, Th. J., Wade Boykin, A., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77–101 [DOI: 10.1037/0003-066X.51.2.77].

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. Los Angeles: Cambridge Assessment/Sage.

Prenzel, M., Walter, O., & Frey, A. (2007). PISA misst Kompetenzen. *Psychologische Rundschau, 58*(2), 128–136 [DOI: 10.1026/0033-3042.58.2.128].

Ramseier, E., & Brühwiler, C. (2003). Herkunft, Leistung und Bildungschancen im gegliederten Bildungssystem. Vertiefte PISA-Analyse unter Einbezug der kognitiven Grundfähigkeiten. *Schweizerische Zeitschrift für Bildungswissenschaften, 25*(1), 23–58.

Rauch, D., & Hartig, J. (2012). Interpretation von Testwerten in der IRT. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 253–264). Berlin: Springer.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? *Psychologische Rundschau, 57*(2), 69–86 [DOI: 10.1026/0033-3042.57.2.69].

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research – Online, 8*, 23–74.

Schladitz, S., Rott, B., Winter, A., Wischgoll, A., Groß Ophoff, J., Hosenfeld, I., Leuders, T., Nückles, M., Renkl, A., Stahl, E., Watermann, R., Wirtz, M., & Wittwer, J. (2013). LeScEd – Learning the Science of Education. Research Competence in Educational Sciences. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.), *The German funding initiative „Modeling and Measuring Competencies in Higher Education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students* (KoKoHs Working Papers No. 3, S. 82–84). Berlin/Mainz: Humboldt-Universität/Johannes Gutenberg-Universität.

Shank, G. D., & Brown, L. (2007). *Exploring educational research literacy*. New York: Routledge.

Spinath, B., & Steinmayr, R. (2012). The roles of competence beliefs and goal orientations for change in intrinsic motivation. *Journal of Educational Psychology, 104*(4), 1135–1148 [DOI: 10.1037/a0028115].

Süß, H.-M. (2003). Intelligenztheorien. In K. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 217–224). Weinheim: Psychologie Verlags Union.

Taub, G. E., Floyd, R. G., Keith, T. Z., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly,* 23(2), 187–198 [DOI: 10.1037/1045-3830.23.2.187].

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

Trapmann, S. (2008). *Mehrdimensionale Studienerfolgsprognose: Die Bedeutung kognitiver, temperamentsbedingter und motivationaler Prädiktoren für verschiedene Kriterien des Studienerfolgs*. Berlin: Logos.

Urban, D., & Mayerl, J. (2008). *Regressionsanalyse: Theorie, Technik und Anwendung*. Berlin/Heidelberg: Springer.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Hrsg.), *IERI monograph series: Issues and methodologies in large scale assessments, Vol. 2* (S. 9–36). Hamburg/Princeton: IERInstitute.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45–65). Seattle: Hogrefe & Huber.

Weiß, R. H. (2006). *CFT 20-R. Grundintelligenztest – Revision*. Göttingen: Hogrefe.

Wigfield, A., & Eccles, J. S. (2000). Expectancy – value theory of motivation. *Contemporary Educational Psychology, 25*, 68–81 [DOI: 10.1006./ceps.1999.1015].

Zlatkin-Troitschanskaia, O., & Kuhn, C. (2010). *Messung akademisch vermittelter Fertigkeiten und Kenntnisse von Studierenden bzw. Hochschulabsolventen – Analyse zum Forschungsstand*. Mainz: Johannes Gutenberg-Universität.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly. An International Journal, 4*(2), 1–19.

Abstract: Being able to make evidence-based decisions is a central aim in higher education. The project LeScEd (Learning the Science of Education) aims at incorporating approaches from fields like Information Sciences, Mathematical Education and Psychology into a comprehensive structure model of Educational Research Literacy (ERL). The current study analyzes the relations of ERL to fluid intelligence and self-reported ERL to analyze discriminant and congruent validity of the developed instrument, respectively. Structural equation modelling revealed small positive effects regarding fluid intelligence and no effects regarding self-reported ERL. These results indicate that ERL is related to, but distinguishable from general intelligence and that self-reported competence is no reliable indicator for actual competence.

Keywords: Educational Research Literacy; Higher Education; Intelligence; Construct Validity; Structural Equation Modeling

Anschrift der Autorinnen/des Autors

Sandra Schladitz, Pädagogische Hochschule Freiburg, Institut für Psychologie,
Kunzenweg 21, 79117 Freiburg, Deutschland
E-Mail: sandra.schladitz@ph-freiburg.de

Dr. Jana Groß Ophoff, Pädagogische Hochschule Freiburg, Institut für Psychologie,
Kunzenweg 21, 79117 Freiburg, Deutschland
E-Mail: jana.grossophoff@ph-freiburg.de

Prof. Dr. Markus Wirtz, Pädagogische Hochschule Freiburg, Institut für Psychologie,
Kunzenweg 21, 79117 Freiburg, Deutschland
E-Mail: markus.wirtz@ph-freiburg.de

Alexandra Winter-Hölzl/Kristin Wäschle/Jörg Wittwer/
Rainer Watermann/Matthias Nückles

Entwicklung und Validierung eines Tests zur Erfassung des Genrewissens Studierender und Promovierender der Bildungswissenschaften

Zusammenfassung: Wissen über das Genre *empirischer Forschungsartikel* ist in den Bildungswissenschaften zentraler Bestandteil der Kompetenz zum wissenschaftlichen Schreiben. Da es bislang keine Verfahren zur Erfassung des Genrewissens gibt, war es unser Ziel, solch einen Test zu entwickeln. *N* = 372 Studierende und Promovierende der Erziehungswissenschaft, Psychologie und des Lehramts beantworteten 37 Genrewissensaufgaben. Diese wurden auf Basis des dichotomen Raschmodells (IRT) skaliert und geeignete Aufgaben ermittelt. 24 Items hielten den Kriterien stand. Die Itemauswahl bildet einen reliablen und ökonomischen Test, der zwischen verschiedenen Studierphasen und Studiengängen differenziert. Der Test könnte Einsicht geben, in welchen Studiengängen Genrewissen erfolgreich vermittelt wird bzw. wo Förderbedarf besteht.

Schlagworte: Genrewissen, Kompetenz, wissenschaftliches Schreiben, Test, IRT

1. Theoretischer Hintergrund

1.1 Wissenschaftliches Schreiben im Studium

Vielen Studierenden bleibt lange unklar, welchen Stellenwert das wissenschaftliche Schreiben im Rahmen ihres Studiums hat und wie bedeutend in diesem Zusammenhang das Wissen über Textsorten bzw. Genres ist. Verstärkt wird dieses Problem durch die mangelnde Vorbereitung an deutschen Gymnasien, insbesondere jedoch durch die meist unsystematische Vermittlung wissenschaftlichen Schreibens an Hochschulen – ein Zustand, der bis in die Qualifikationsphase Promovierender anhält (Kamler & Thomson, 2006). Im Rahmen informeller Kommunikationssituationen zwischen Studierenden und Dozierenden bleibt das Wissen über Schreibprozesse und Textprodukte meist implizit (Schindler, 2008). Die Folgen aufseiten der Studierenden zeigen sich beispielsweise in einer von Dittmann, Geneuss, Nennstiel und Quast (2003) durchgeführten Befragung, in der mehr als drei Viertel einer Stichprobe von 283 Studierenden angaben, Schreibprobleme zu haben und sich damit „allein gelassen“ zu fühlen (51.3 %).

Im Rahmen wissenschaftlicher Schreibsozialisation sind Schreibprobleme jedoch eine besondere Herausforderung, weil wissenschaftliche Kommunikation neben einer Ausbildungsfunktion in der Hochschullehre und der Verbreitung von Fachwissen in erster Linie den Zweck erfüllen soll, einen gleichermaßen effektiven wie effizienten

Austausch unter Wissenschaftlerinnen zu ermöglichen (Goldman & Bisanz, 2002). Und genau dieser konfrontiert Studierende beim akademischen Schreiben mit erhöhten Anforderungen an die Textsortenkompetenz (d. h. Kenntnissen über unterschiedliche Funktionen und Formen beispielsweise einer Zusammenfassung gegenüber einer Hausarbeit), die eng mit Stil- und rhetorischer Kompetenz verknüpft ist (Kruse, Jakobs & Ruhmann, 1999). Denn so verschieden die Meinungen sind, was unter ‚gutem' wissenschaftlichem Schreiben zu verstehen ist, so disziplinabhängig sind auch die Formate wissenschaftlichen Schreibens: Sie folgen Regeln, die sich je nach Fachbereich in unterschiedlichen Textsorten oder Genres manifestieren.

1.2 Formate wissenschaftlichen Schreibens: Das Genre

Definition und Funktion des Genres

Genre bezeichnet eine Klasse kommunikativer Abläufe, die Mitgliedern einer Wissens- oder Diskursgemeinschaft (wie z. B. einer akademischen Forschergruppe) zur Erreichung rhetorischer Ziele dienen. Genres sollen mittels ihrer Konventionen dazu befähigen, genau, wirkungsvoll und effektiv Wissen zu kommunizieren (Bazerman, 1983; Rijlaarsdam, Couzijn, Janssen, Braaksma & Kieft, 2006; Swales, 1990).

Dazu bietet die Kenntnis über das jeweilige Genre relevante Anhaltspunkte zur Gliederung eines Textes und zur Form und Funktion seiner Abschnitte. Schreibende können sich auf diese Weise an einer Art Konstruktionsanleitung für das Textprodukt orientieren (Goldman & Bisanz, 2002). Daher ist bei neuen Mitgliedern einer Diskursgemeinschaft wie z. B. Doktorandinnen das Wissen über Genres ein wichtiges Lernziel im Rahmen der Schreibsozialisation. Fundiertes Genrewissen ist eine wesentliche Grundlage für die zielgerichtete Planung und Ausführung des Schreibprozesses (Kruse, 1997).

Der empirische Forschungsartikel

Im Bemühen der Diskursgemeinschaft, Qualität und Vergleichbarkeit zu gewährleisten, wurde der Forschungsartikel insbesondere im Bereich der Psychologie und in weiten Teilen der „Educational Sciences" zu einem Schlüsselgenre im internationalen und nationalen Diskurs (Swales, 1990). Nach Swales ist der Forschungsartikel deshalb ein Schlüsselgenre, weil er quasi im Zentrum unterschiedlicher Genres steht, die im Forschungsprozess Verwendung finden. Beispielsweise folgen Masterarbeiten oder Dissertationen dessen Aufbau. Zugleich weist er eine hohe Verwendungshäufigkeit und Veröffentlichungspraxis in wissenschaftlichen Zeitschriften und Journals auf. Da sich die Bildungswissenschaften heutzutage in weiten Teilen als empirische Disziplinen begreifen, ist der empirische Forschungsartikel auch hier entsprechend häufig verbreitet und Wissenschaftlerinnen in diesem Bereich müssen sich demnach in hohem Maße an die entsprechenden rhetorischen Konventionen (APA, 2009; DGPs-Richtlinien, 2007) halten, um in ihrer Disziplin erfolgreich sein zu können (Gruber et al., 2006).

1.3 Wissen über das Genre „empirischer Forschungsartikel" als Determinante wissenschaftlicher Schreibkompetenz

Aus kognitionspsychologischer Perspektive sind deklaratives und prozedurales Wissen zentrale Bedingungen kompetenten Handelns (Simonton, 2003). Entsprechend kann Genrewissen als zentrale Voraussetzung wissenschaftlicher Schreibkompetenz aufgefasst werden. In Andersons ACT-R-Theorie (Anderson, Matessa & Lebiere, 1997) wird Kompetenzerwerb als Aufbau und sukzessive Prozeduralisierung deklarativer Wissensschemata modelliert. Beispielsweise erkennen Schreibende im Sinne des Erwerbs deklarativen Genrewissens, dass eine Einleitung unter anderem zur Darstellung der Problemrelevanz dient oder im Diskussionsteil Einschränkungen der Studie genannt werden sollten. Erst durch wiederholtes Üben können anschließend zunächst getrennte Wissensschemata miteinander verknüpft werden.

Die Rolle des Genrewissens beim Schreiben kann mithilfe der psychologischen Schreibforschung näher spezifiziert werden. Hayes und Flower (1980) beschreiben in ihrem Modell unterschiedliche kognitive Prozesse und mentale Repräsentationen im Schreibprozess. Durch ihre theoretische Annahme, einen Schreibauftrag als schlecht definiertes Problem innerhalb eines *semantischen* und eines *rhetorischen* Problemraums zu sehen, kann Schreiben als Problemlöseprozess konzeptualisiert werden. Dieser verläuft allerdings nicht linear, sondern vielmehr zyklisch und interaktiv über den wechselseitigen Abruf von Zielen, mentalen Repräsentationen und kognitiven Prozessen aufseiten der Textproduzenten. Dabei wird im Rahmen eines Schreibprozesses relevante Information aus dem Langzeitgedächtnis generiert, organisiert und die Zielsetzung definiert *(planning),* inhaltliches Wissen in geschriebene Sprache überführt *(translating)* und in einem Überarbeitungsprozess die Verbesserung der Textqualität angestrebt *(reviewing)*.

In Bezug auf den rhetorischen Problemraum erachten wir deshalb in Anlehnung an Hayes und Flower (1980) drei unterschiedliche Wissensformen als wesentlich: (1) deklaratives rhetorisches Wissen, (2) diagnostisches rhetorisches Wissen sowie (3) rhetorisches Wissen in der Textproduktion. Während das deklarative rhetorische Wissen das aktiv verfügbare und reproduzierbare Wissen über rhetorische Prinzipien und Konventionen meint, geht es beim diagnostischen rhetorischen Wissen darum, deklaratives rhetorisches Wissen im Sinne einer Bewertungsfähigkeit auf Schreibprodukte anzuwenden – also beispielsweise ein ‚gutes' Abstract von einem ‚schlechten' Abstract unterscheiden zu können. In Hinblick auf das rhetorische Wissen in der Textproduktion sollten Schreibende schließlich in der Lage sein, rhetorische Prinzipien in der Textproduktion flexibel und situationsangemessen anwenden zu können, d. h. für Dritte flüssig lesbare wissenschaftliche Texte verfassen können. Die Erfassung der letztgenannten Wissensart in Form einer Textproduktionsaufgabe ist jedoch so aufwendig, dass sie in Tabelle 1 aus Gründen der Vollständigkeit zwar erläutert, aber in einer eigenen Studie weiter untersucht wird. In diesem Text beschränken wir uns auf die psychometrische Analyse der ersten beiden Wissensformen deklaratives und diagnostisches rhetorisches Wissen.

Die in Hinblick auf das Genre *empirischer Forschungsartikel* relevanten rhetorischen Prinzipien und Konventionen haben wir durch Sichtung der Richtlinien und Manuale einschlägiger wissenschaftlicher Vereinigungen (APA, 2009; DGPs-Richtlinien, 2007) sowie verbreiteter Ratgeberliteratur (Silvia, 2007; Strunk & White, 1999) und publizierter Artikel zum wissenschaftlichen Schreiben (Bem, 2003) zusammengestellt. Sie lassen sich als inhaltliche *Wissensanforderungen* grob kategorisieren als (a) rhetorisches Wissen zu Hauptabschnitten des empirischen Forschungsartikels – bestehend aus Einleitung, Methode, Ergebnisteil und Diskussion –, (b) rhetorisches Wissen zu Nebenabschnitten – bestehend aus Titel, Abstract, Literatur und Anhang – sowie (c) Wissen über wissenschaftlichen Schreibstil. Letztgenannte Anforderung umfasst Stilmerkmale wie Prägnanz und Präzision, die z. B. durch die Vermeidung von Füllwörtern, Redundanzen, Passivkonstruktionen, Substantivierungen usw. realisiert werden. Die Kategorisierung der Anforderungen ergibt sich hauptsächlich aus den Hinweisen zu Struktur und Inhalt eines Forschungsartikels (APA, Kapitel 2, *Manuscript Structure and Content*) und zum bevorzugten Schreibstil (APA, Kapitel 3, *Writing Clearly and Consisely*), aber auch über die von Swales empirisch belegte Einteilung der Hauptstruktur eines Forschungsartikels und seiner Nebenabschnitte. So ergibt sich zur Modellierung von Genrewissen im Bereich empirischer Forschungsartikel das in Tabelle 1 dargestellte Raster in Hinblick auf die systematische Entwicklung unserer Testaufgaben.

Auf Basis dieser Modellierung möchten wir prüfen, inwieweit in unserer Stichprobe (1) deklaratives rhetorisches Wissen im Sinne von Wissen über Konzepte und Bestandteile von Konzepten (z. B. welche Inhalte in einem Abstract formuliert sein sollten) und (2) diagnostisches rhetorisches Wissen im Sinne einer für spätere Revisionsprozesse bedeutsamen Bewertungsfähigkeit vorhanden ist (z. B. „Was ist an dieser Darstellung zu kritisieren?“). Wie sich die Umsetzung der ersten Wissensformen in der Wissensform (3) rhetorisches Wissen in der Textproduktion niederschlägt, wird in einer separaten Studie u. a. anhand dieser Beispielaufgabe erforscht: „Nutzen Sie bitte die Inhalte des folgenden Dialogs, um ein Abstract zu formulieren, das am Beginn eines empirischen Forschungsartikels stehen könnte.“

	Inhaltliche Wissensanforderungen		
Instruktionsbeispiele je Wissensform	**Hauptabschnitte**	**Nebenabschnitte**	**wiss. Stil**
Deklaratives rhetorisches Wissen Was ist…? Wozu dient…? Welche Bestandteile hat…?	13 Items	5 Items	3 Items
Diagnostisches rhetorisches Wissen Was ist an dieser Darstellung (…) zu kritisieren…? Was fehlt (…)? Was sollte (…) präziser formuliert werden?	7 Items	3 Items	6 Items
Rhetorisches Wissen in der Textproduktion Bitte schreiben Sie nun (…) ein Abstract (…).	2 Aufgaben (Schreiben)		

Tab.1: Modell zu Genrewissen im Rahmen der Aufgabenkonstruktion

In allen bildungswissenschaftlichen Studiengängen sollen Studierende und Promovierende beim Erlernen wissenschaftlicher Schreibkompetenz rhetorisches Wissen in Form von Genrewissen erwerben, doch ist bislang wenig bekannt, in welchem Ausmaß sie das tun. Vor diesem Hintergrund bestand unser Ziel darin, ein reliables, valides und testökonomisches Instrument zur Erfassung des Wissens von Studierenden und Promovierenden der Bildungswissenschaften über das Genre *empirischer Forschungsartikel* zu entwickeln. Da es bislang keine standardisierten Tests zur Erfassung von Genrewissen gibt, könnte unser Test Einsicht bieten, in welchen Studiengängen eine erfolgreiche Vermittlung von Genrewissen geleistet wird – oder in welchen Studiengängen Förderbedarf besteht.

2. Testentwicklung

2.1 Vorstudien

Zur Realisierung eines breiten Aufgabenspektrums wurden über 100 Aufgaben als Mehrfachauswahlaufgaben, Kurzantworten und offene Antworten entwickelt. Mithilfe des Retrospective-Think-Aloud-Verfahrens, bei dem die Teilnehmenden erst eine Aufgabe abschließen, bevor sie ihre Gedankengänge retrospektiv artikulieren, wurden die Aufgaben in einer Vorstudie ($N = 7$) auf Verständlichkeit hin überprüft.

Auf Basis dieser Ergebnisse wurden die Aufgabeninhalte von fünf Expertinnen und Experten der Bildungswissenschaften beurteilt, die sich sowohl durch renommierte Publikationen als auch durch ein breites Erfahrungsspektrum in der Betreuung von Schreibprozessen von Doktorandinnen und Habilitandinnen auszeichneten.

Nach einer darauffolgenden Pilotstudie mit Studierenden ($N = 82$) und Promovierenden ($N = 5$) der Bildungswissenschaften wurden die Items weiter optimiert. Eine Aufgabe zur Textproduktion wurde aufgrund der zu aufwendigen Bearbeitungszeit aus dem Itempool der vorliegenden Studie herausgenommen. Diese Schreibaufgabe kommt derzeit in einer Folgestudie zur Vorhersage der Qualität wissenschaftlicher Schreibprodukte auf Basis des Genrewissens zur Anwendung.

2.2 Studie zu Genrewissen und wissenschaftlichem Schreiben

Nach einer letzten Auswahl und Optimierung der Aufgaben durch eine Expertenrunde ($N = 5$) wurden 37 Items in den Test aufgenommen und parallel von drei weiteren Experten hinsichtlich ihrer Inhaltsvalidität eingeschätzt. Zur Veranschaulichung dazu dienen zwei Itembeispiele in Abbildung 1.

Diagnostisches rhetorisches Wissen zu Hauptabschnitten

27. Lesen Sie bitte diesen Ausschnitt aus dem *Methodenteil* eines empirischen Forschungsartikels:

[…] *Probandinnen und Probanden.* An der Untersuchung nahmen ca. 80 Psychologie-Studierende der Albin-Gregor-Universität gegen Bezahlung (10 € pro Versuchsteilnahme) teil. Das durchschnittliche Alter der Probandinnen und Probanden lag bei 23.5 Jahren (*SD* = 4.6 Jahre) […]

Was ist an dieser Darstellung vorrangig zu kritisieren? (Eine Antwort ist richtig)

☐ Bei den Altersangaben fehlt die Spannweite (18 bis 27 Jahre).
☒ Die Angabe zur Anzahl der Teilnehmenden ist ungenau.
☐ Studierende dürfen für eine Versuchsteilnahme nicht entschädigt werden.
☐ Standardabweichungen (*SD*) werden im Ergebnisteil berichtet.

Deklaratives rhetorisches Wissen zu Nebenabschnitten

15. Nach dem Titel folgt das Abstract (*dt.:* Zusammenfassung). Bitte nennen Sie stichwortartig „typische" Inhalte, die im Abstract eines empirischen Forschungsartikels formuliert werden sollten:

……

[Forschungsfeld – Forschungsfrage – Stichprobe – Stichprobengröße – Methode – Hauptergebnis – Schlussfolgerung]

Abb. 1: Itembeispiele und deren Lösungen

2.3 Kodierung

Die Kodierung der Antworten zu den Items (kein Punkt = 0 oder ein Punkt = 1) orientierte sich überwiegend an den Konventionen des Publication Manuals der APA (American Psychological Association, 2009) und zusätzlichen Experteneinschätzungen. Items im offenen Antwortformat wurden anhand eines differenzierten Antwortkatalogs für korrekte Begriffe sowie deren gültiger Beschreibung ebenfalls mit null Punkten bzw. einem Punkt bewertet. Neben den Richtlinien der APA diente auch die Veröffentlichungspraxis deutschsprachiger wissenschaftlicher Zeitschriften als Wertungsgrundlage (z. B. Zeitschrift für Psychologie). So wurden der gesuchte Begriff ‚Theorie' wie auch dessen treffende Umschreibungen – wie z. B. *Theorieteil* – als richtig bewertet, aber auch Begriffe wie *Einführung* oder *Einleitung,* da diese in wissenschaftlichen Publikationen auf der gleichen Ebene verwendet werden. Für die Antwort *Hauptteil* wurde dagegen kein Punkt vergeben.

3. Methode

3.1 Stichprobe

An der Studie nahmen $N = 300$ Studierende ($M = 23.2$ Jahre, $SD = 4.4$) und $N = 72$ Promovierende ($M = 29.3$ Jahre, $SD = 4.3$) der bildungswissenschaftlichen Studiengänge Lehramt ($N = 266$), Erziehungswissenschaft ($N = 37$) und Psychologie ($N = 69$) teil. Dabei waren 49 der 72 Promovierenden Absolventinnen und Absolventen der Psychologie, 20 Promovierende der Erziehungswissenschaft sowie drei Promovierende aus dem Lehramtsstudiengang. Als Belohnung für die Teilnahme konnten Preise (z. B. Gutscheine) im Rahmen einer Verlosung gewonnen werden. Die Studierenden befanden sich zwischen dem ersten bis achten Semester; über zwei Drittel der Teilnehmenden waren Frauen.

3.2 Datenerhebung

Im Rahmen von Lehrveranstaltungen erhielten die Teilnehmenden die 37 Testfragen sowie fünf weitere Aufgaben zu Statistical und Informational Literacy unseres Verbundpartners als Papierfragebogen. Die Bearbeitung dauerte zwischen 50 und 70 Minuten. Die Bearbeitungszeit war nicht begrenzt, um die Testschwierigkeit für Studierende mit erst wenigen Semestern Studium nicht zusätzlich zu erhöhen.

3.3 Instrumente

Test zu Genrewissen
Der Test bestand aus neun Items zum wissenschaftlichen Schreibstil, 20 Items zu Hauptabschnitten sowie acht Items zu Nebenabschnitten (Itembeispiele in Abb. 1), wovon 21 Items der deklarativen und 16 Items der diagnostischen Wissensform zuzuordnen sind. Alle offenen Antworten wurden von zwei unabhängigen Ratern komplett kodiert. Die Interrater-Reliabilität wurde mittels Cohens Kappa (Wirtz & Caspar, 2002) als zufallskorrigiertes Maß berechnet. Die Rater-Übereinstimmung lag bei allen 14 offenen Items mit $\kappa = .91$ bis $\kappa = 1.0$ im sehr guten Bereich. Fehlende Werte, die es durch die ausreichend gegebene Bearbeitungszeit und den Anreiz der Verlosung fast nur bei offenen Antworten gab (2 % bis 6.5 %), wurden dabei als falsch kodiert.

Validierungsvariablen
Zu Testbeginn wurden zur späteren Validierung der abhängigen Variablen *Genrewissen* neben soziodemografischen Daten die Variablen Studiengang – anhand der drei Kategorien Lehramt, Erziehungswissenschaft und Psychologie –, die Studierphase – anhand der drei Kategorien 1. bis 4. Semester, 5. Semester bis Master und Promovierende – sowie die Deutschnote miterhoben. Zusätzlich wurden als mögliche Lerngelegenheiten

die Häufigkeit des Lesens von e-papers (*Wie oft lesen Sie e-papers?* mit den Antwortoptionen sehr selten oder nie, einmal im Monat, einmal in der Woche, mehrmals in der Woche, täglich, mehrmals täglich), das (Mit-)Verfassen eines Forschungsartikels (*Haben Sie bereits einen Forschungsartikel, der in einem wissenschaftlichen Journal veröffentlicht wurde/wird, selbst verfasst oder daran mitgeschrieben?* mit der Antwortoption ja, nein) sowie das Anfertigen von Forschungsarbeiten (*Haben Sie im Laufe Ihres Studiums bereits eine oder mehrere der folgenden Arbeiten angefertigt?* mit maximal fünf Punkten, wenn folgende Arbeiten genannt wurden: Hausarbeit, Bachelorarbeit, Masterarbeit, wissenschaftliches Poster, sonstige Forschungsarbeiten) erfragt.

3.4 Psychometrische Analyse

Unter der Grundannahme eines eindimensionalen Konstrukts wurde der Test auf Basis des dichotomen Raschmodells (Rasch, 1960; Rost, 2004) mit dem Programm ConQuest (Wu, Adams & Wilson, 1998) skaliert. Die Raschhomogenität der Items wurde anhand von Kennwerten (Bond & Fox, 2007; Adams, 2002) überprüft. Die Prüfung der Eindimensionalität des Genrewissenstests erfolgte mit Mplus (Muthén & Muthén, 2001). Zur Überprüfung der Konstruktvalidität des Tests wurde in ConQuest ein latentes Regressionsmodell mit den genannten unabhängigen Variablen spezifiziert.

4. Ergebnisse

4.1 Psychometrische Qualität

Anhand von Tabelle 2 werden die teststatistischen Kennwerte aller 37 Items berichtet.

Estimates sind die im Raschmodell ermittelten Itemschwierigkeiten. Die *Weighted Mean Squares* mit der zugehörigen *Spanne* (CI) erlauben Aussagen über Abweichungen von empirischen und erwarteten Werten und sind ein Itemfitmaß mit einem Erwartungswert von 1. Ebenfalls stellen *T-Werte* sowohl über +2 als auch unter −2 eine signifikante Abweichung von diesem Erwartungswert dar (Wu, Adams, Wilson & Haldane, 2007). Dabei deuten T-Werte über +2 auf sogenannte *Misfits* hin, die wenig zur Schätzung der latenten Variable beitragen und als problematisch gelten (Wilson, 2005). Während Items mit T-Werten über +2 verworfen wurden, akzeptierten wir Items unter einem Wert von −2, da diese sogenannten *Overfits* in der Regel als unproblematisch (Jude, 2006) gelten. Overfits deuten auf eine höhere Trennschärfe hin. Die höher ausfallenden Trennschärfen bei den Items 1 bis 5, 7 bis 11 sowie 26 und 31 weisen demnach darauf hin, dass die genannten Items besonders gut zwischen der latenten Fähigkeit differenzieren. Entsprechend wurden alle Items mit einem wMNSQ-Wert deutlich kleiner als 1.0 akzeptiert.

Unter der beschriebenen Beachtung der Itemfit-Kennwerte (wMNSQ-Spannen, *T*-Werte), der Itemschwierigkeiten und der Itemtrennschärfeparameter erwiesen sich 24 Items als geeignet – davon zehn im geschlossenen Antwortformat. Die ermittelte Va-

Item	Wissens-form	Anforderung	Estimate	wMNSQ	CI	*T*	p_i	r_{it}	Passung
20	diagnos-tisch (dia)	wiss. Stil (WS)	−3.320	1.06	[0.63, 1.37]	0.4	0.94	0.15	nein
34	deklarativ (dek)	Nebenab-schnitt (NA)	−3.178	1.02	[0.66, 1.34]	0.2	0.93	0.17	nein
12	dek	NA	−1.973	1.00	[0.84, 1.16]	0.0	0.81	0.31	ja
25	dek	Hauptab-schnitt (HA)	−1.598	1.10	[0.88, 1.12]	1.5	0.76	0.23	nein
37	dek	S	−0.949	1.26	[0.92, 1.08]	5.7	0.64	0.09	nein
27	dia	HA	−0.922	0.95	[0.92, 1.08]	−1.1	0.63	0.42	ja
32	dek	HA	−0.774	1.04	[0.92, 1.08]	1.1	0.60	0.38	ja
6	dia	HA	−0.696	0.94	[0.92, 1.08]	−1.7	0.59	0.48	ja
33	dek	HA	−0.669	1.05	[0.93, 1.07]	1.3	0.58	0.39	ja
14	dia	NA	−0.657	0.98	[0.93, 1.07]	−0.4	0.58	0.44	ja
35	dia	NA	−0.526	1.09	[0.93, 1.07]	2.4	0.55	0.32	nein
24	dek	HA	−0.384	1.06	[0.93, 1.07]	1.7	0.52	0.33	ja
22	dia	WS	−0.241	1.17	[0.92, 1.08]	4.2	0.49	0.23	nein
21	dia	WS	−0.229	1.16	[0.92, 1.08]	3.8	0.49	0.27	nein
28	dek	HA	−0.163	1.29	[0.84, 1.16]	6.6	0.48	0.1	nein
29	dek	HA	0.022	0.96	[0.93, 1.07]	−0.8	0.44	0.45	ja
3	dek	HA	0.041	0.83	[0.91, 1.09]	−4.1	0.43	0.59	ja
18	dia	WS	0.209	1.27	[0.91, 1.09]	5.3	0.40	0.14	nein
9	dia	HA	0.210	0.73	[0.91, 1.09]	−6.3	0.40	0.71	ja
30	dia	HA	0.336	1.04	[0.66, 1.34]	0.9	0.38	0.40	ja
15	dek	NA	0.448	0.92	[0.90, 1.10]	−1.5	0.35	0.53	ja
2	dek	HA	0.500	0.69	[0.89, 1.11]	−6.4	0.35	0.74	ja
16	dia	WS	0.535	1.21	[0.89, 1.11]	3.6	0.34	0.19	rev
13	dek	NA	0.550	1.35	[0.89, 1.11]	5.7	0.34	0.07	nein
19	dia	WS	0.639	1.26	[0.89, 1.11]	4.2	0.32	0.18	nein
11	dek	WS	0.716	0.73	[0.88, 1.12]	−4.9	0.31	0.71	ja
5	dia	HA	0.733	0.70	[0.88, 1.12]	−5.5	0.30	0.74	ja
26	dek	HA	0.733	0.78	[0.88, 1.12]	−4.0	0.30	0.66	ja
8	dia	HA	0.763	0.88	[0.88, 1.12]	−2.0	0.30	0.55	ja
31	dek	HA	0.860	0.63	[0.88, 1.12]	−6.6	0.28	0.80	ja
7	dia	HA	0.925	0.65	[0.87, 1.13]	−6.1	0.27	0.79	ja
10	dek	HA	1.026	0.67	[0.87, 1.13]	−5.5	0.26	0.77	ja
1	dek	HA	1.162	0.59	[0.86, 1.14]	−6.6	0.24	0.84	ja
17	dia	NA	1.435	1.21	[0.84, 1.16]	2.4	0.20	0.12	rev
36	dek	WS	1.459	1.03	[0.84, 1.16]	0.4	0.19	0.36	ja
23	dek	NA	1.478	1.05	[0.84, 1.16]	0.6	0.19	0.37	ja
4	dek	HA	1.499	0.75	[0.83, 1.17]	−3.2	0.19	0.67	ja

Legende. dek = deklaratives rhetorisches Wissen; dia = diagnostisches rhetorisches Wissen; HA = Hauptabschnitt (Einleitung, Methode, Ergebnis, Diskussion); NA = Nebenabschnitt (Titel, Abstract, Literatur); WS = wissenschaftlicher Schreibstil; Estimate = Itemschwierigkeit (Raschmodell); wMNSQ = weighted Mean Square; CI = wMNSQ-Spanne; *T* = Wert aus *T*-Verteilung; p_i = Itemschwierigkeit (Klassische Testtheorie); r_{it} = Trennschärfe (Klassische Testtheorie); Passung = ja, nein, rev = revidiert (Item ging in überarbeiteter Form in eine neue Erhebung ein).

Tab. 2: Psychometrische Kennwerte der Aufgaben und ihre Zuordnung nach Wissensform und Anforderungsniveau (Items nach Schwierigkeit absteigend sortiert)

rianz für die Personenparameter beträgt dabei 0.989 und die EAP/PV-Reliabilität liegt bei 0.897 (welche beim eindimensionalen Raschmodell mit Cronbach's α vergleichbar ist; vgl. Bond & Fox, 2007; Rost, 2004).

Die Ergebnisse einer konfirmatorischen Faktorenanalyse in Mplus zur Überprüfung der Modellgüte (24 Items) deuteten auf eine zufriedenstellende Lösung bezüglich des eindimensionalen Modells hin: Der χ^2-Test wurde zwar signifikant ($\chi^2 = 342.33$, $df = 252$, $p < 0.1$), der Quotient aus χ^2 und Freiheitsgraden blieb jedoch deutlich unter 2. Zudem wiesen der RMSEA = 0.031, der CFI = 0.996 und der TLI = 0.996 auf eine sehr gute Passung des Modells hin (Backhaus, Erichson & Weiber, 2011; Bühner, 2006). Demgegenüber zeigten sich bei einem möglichen zweidimensionalen Modell ($\chi^2 = 340.41$, $df = 251$, $p < 0.1$), das die beiden Wissensformen deklaratives und diagnostisches rhetorisches Wissen unterscheidet (RMSEA = 0.031, CFI = 0.996, TLI = 0.996), bzw. bei einem dreidimensionalen Modell ($\chi^2 = 341.23$, $df = 249$, $p < 0.1$), bei dem die inhaltlichen Anforderungen Hauptabschnitt, Nebenabschnitt und Schreibstil als mögliche Dimensionen in Betracht kommen (RMSEA = 0.032, CFI = 0.996, TLI = 0.996), keine verbesserten Fitmaße. Das eindimensionale Modell erwies sich demzufolge als nicht signifikant schlechter als die mehrdimensionalen Modelle.

In einer sog. *Wright Map* werden zur Veranschaulichung die Itemschwierigkeiten und Personenfähigkeiten dargestellt, die näherungsweise normalverteilt waren (Kolmogorov-Smirnov-Z = 1.04, $p = .23$, *ns;* s. Abb. 2).

Unter den 24 Items verblieben somit elf Items zu deklarativem rhetorischen Wissen über Hauptabschnitte (Einleitung, Methode, Ergebnis, Diskussion) und drei Items über Nebenabschnitte (Titel, Abstract), zwei Items zu wissenschaftlichem Schreibstil, sieben Items zu diagnostischem rhetorischen Wissen über Hauptabschnitte (Einleitung, Methode, Ergebnis, Diskussion) sowie ein Item zu diagnostischem rhetorischen Wissen über Nebenabschnitte (Titel) mit Cronbach's α = .91. Inhaltlich zählten dabei erwartungsgemäß Fragen zum wissenschaftlichen Schreibstil (siehe Item 36, Abb. 2) zu den schwierigeren Items, wohingegen Fragen wie z. B. das Aufzählen der Inhalte eines Nebenabschnitts (siehe Item 15, Abb. 2) kennzeichnend für ein mittleres Schwierigkeitsniveau waren und Aufgaben beispielsweise zur Diagnose sprachlicher Ungenauigkeiten in einem Hauptabschnitt (siehe Item 27, Abb. 2) von den meisten Befragten beantwortet werden konnten.

4.2 Konstruktvalidität: Korrelationen mit Außenkriterien

Um einen ersten deskriptiven Eindruck bezüglich der Konstruktvalidität zu erhalten, werden in Tabelle 3 die bivariaten Zusammenhänge der Analysevariablen dargestellt. Wir erwarteten, dass die Variable Studiengang insofern einen Einfluss auf das Testergebnis aufweisen würde, als dass das Psychologiestudium mehr Lerngelegenheiten und einen aktiveren Umgang mit dem Genre *empirischer Forschungsartikel* bietet als z. B. das Lehramtsstudium. Zudem sollten das Lesen von e-papers, das (Mit-)Verfassen eines Forschungsartikels sowie das Anfertigen von Forschungsarbeiten als Lerngelegenhei-

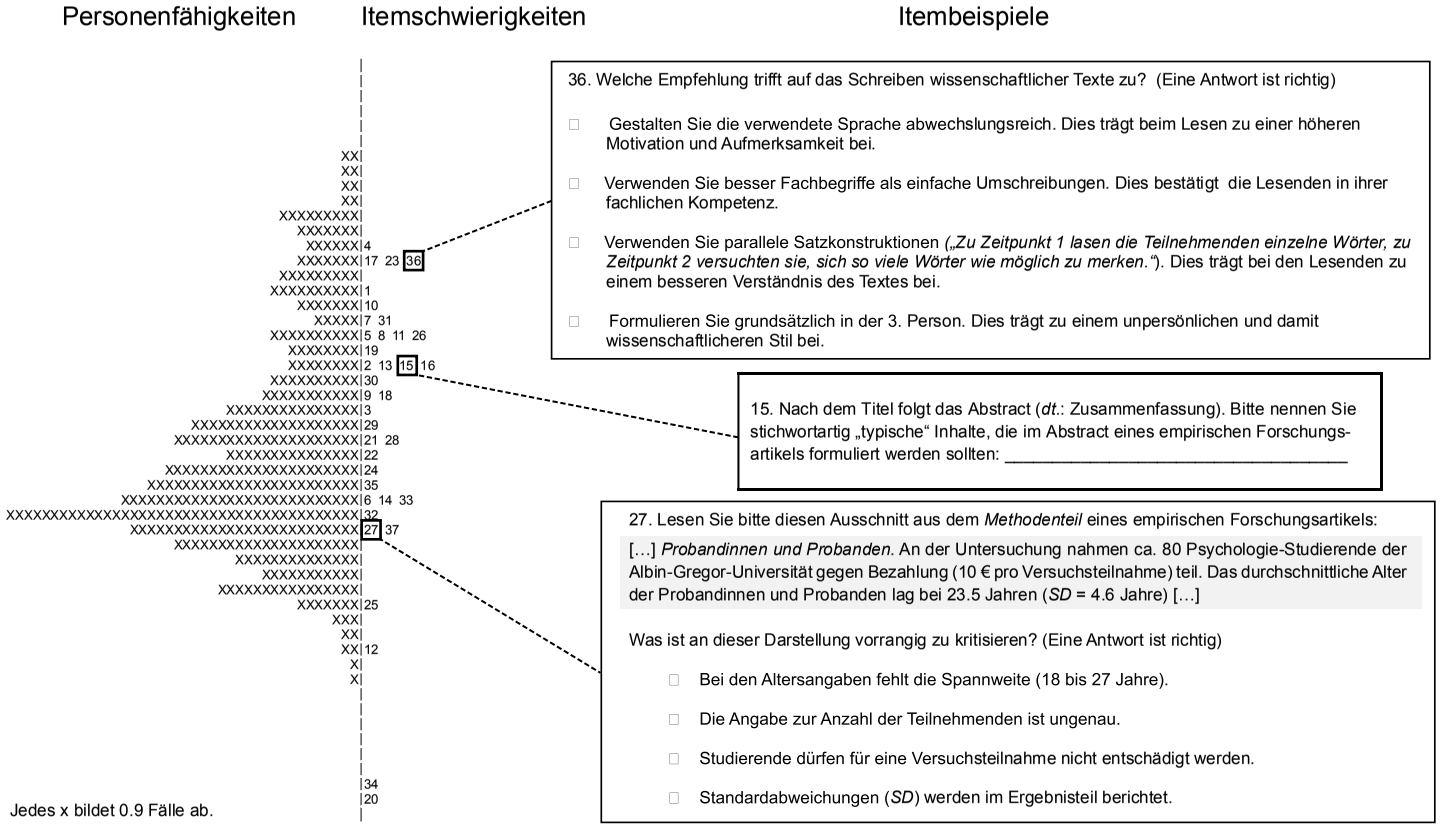


Abb. 2: Darstellung der Personenfähigkeiten und Itemschwierigkeiten des Genrewissenstests in Form einer Wright Map. Die Itembeispiele 15 und 27 finden sich bereits in Abbildung 1.

ten für den Erwerb von Genrewissen positiv mit dem erreichten Testwert korreliert sein, auch wenn Forschungsarbeiten stark an die Variable Studierphase gekoppelt sind. Die Deutschnote dagegen sollte mit Genrewissen wenig in Verbindung stehen, da dies sonst ein Hinweis darauf sein könnte, dass die Items z. B. insbesondere das Leseverständnis von Befragten testen.

Erwartungsgemäß korrelierte die Deutschnote – dabei gilt: je besser das Testergebnis, desto niedriger (d. h. besser) die Deutschnote – nur geringfügig mit der Testleistung ($r = -.17$, $p < .01$). Das Lesen von e-papers ($r = .37$, $p < .01$) sowie das Verfassen von Forschungsartikeln ($r_S = .44$, $p < .01$) oder Forschungsarbeiten ($r_S = .63$, $p < .01$) wiesen hingegen einen höheren Zusammenhang mit Genrewissen auf. Weiter zeigte sich, dass der Studiengang nicht nur mit dem Testergebnis, sondern auch mit den eben genannten Variablen des Lesens von e-papers ($r_S = .37$, $p < .01$), des Anfertigens von Forschungsartikeln ($r_S = .41$, $p < .01$) oder Forschungsarbeiten ($r_S = .64$, $p < .01$) zusammenhing. Dabei korrelierten der Studiengang und die Studierphase ($r_S = .63$, $p < .01$), welche wiederum mit dem Anfertigen von Forschungsarbeiten verknüpft ist ($r_S = .71$, $p < .01$).

Mittels einer latenten Regressionsanalyse in ConQuest analysierten wir, wie die Zusammenhänge der beschriebenen Variablen mit dem Testergebnis ausfallen, wenn man jeweils für den Einfluss der anderen Variablen kontrolliert. Da in unserer Stichprobe die Variablen Studiengang und Studierphase konfundiert waren (Tab. 3), werden die Ergeb-

Variable	Studiengang	Deutschnote	Lesen e-papers	Forschungsartikel	Studierphase	Forschungsarbeiten	Genrewissen
Studiengang		$-.07_S$	$.37_S$**	$.41_S$**	$.63_S$**	$.64_S$**	$.69_S$**
Deutschnote			−.02	$-.14_S$**	$-.16_S$**	$-.07_S$**	−.17**
Lesen e-papers				$-.23_S$**	$.34_S$**	$.34_S$**	.37**
Forschungsartikel					$.54_S$**	$.47_S$**	$.44_S$**
Studierphase						$.71_S$**	$.63_S$**
Forschungsarbeiten							$.63_S$**
Genrewissen							

Legende. Studiengang = Lehramt (−1), Erziehungswissenschaft (0), Psychologie (+1); Deutschnote: 1 = sehr gut, 2 = gut, 3 = befriedigend, 4 = ausreichend; Lesen e-papers = *Wie oft lesen Sie e-papers?* (1 = sehr selten oder nie, 2 = einmal im Monat, 3 = einmal in der Woche, 4 = mehrmals in der Woche, 5 = täglich, 6 = mehrmals täglich); Forschungsartikel = *Haben Sie bereits einen Forschungsartikel, der in einem wissenschaftlichen Journal veröffentlicht wurde/wird, selbst verfasst oder daran mitgeschrieben?* (nein = 1, ja = 2); Studierphase: Semester 1 bis 4 = −1, Semester 5 und höher bzw. Master = 0, Promovierende = +1; Forschungsarbeiten = *Haben Sie im Laufe Ihres Studiums bereits eine oder mehrere der folgenden Arbeiten angefertigt?* (maximal 5 Punkte, wenn folgende Arbeiten genannt wurden: Hausarbeit, Bachelorarbeit, Masterarbeit, wissenschaftliches Poster, sonstige Forschungsarbeiten); Genrewissen = Testwert (24 Items).

$_S$ = Spearman; * $p < 0.05$; ** $p < 0.01$.

Tab. 3: Interkorrelationsmatrix zwischen der Kriteriumsvariablen Testergebnis und den Validierungsvariablen

	Modell 1: alle Variablen			Modell 2: ohne Variable *Studiengang*			Modell 3 ohne Variable *Studierphase*		
	unstandardisiert		standardisiert	unstandardisiert		standardisiert	unstandardisiert		standardisiert
Variable	**Koeffizient**	***SE***	**Koeffizient**	**Koeffizient**	***SE***	**Koeffizient**	**Koeffizient**	***SE***	**Koeffizient**
Konstante	0.047	0.110		−0.729	0.121		−0.086	0.110	
Lehramt	−0.698**	0.077	−.49				−0.746**	0.081	−.55
Psychologie	0.860**	0.073	.57				0.832**	0.081	.61
Deutschnote	−0.097**	0.025	−.02	−0.098*	0.042	−.03	−0.105**	−0.03	−.03
e-papers	0.038*	0.016	.15	0.086**	0.025	.02	0.054**	0.01	.01
Fo-Artikel	0.321**	0.066	.19	0.202	0.109	.18	0.348**	0.068	.22
Semester 1–4	−0.155**	0.049	−.07	−0.201*	0.080	−.13			
Promovierende	0.116	0.076	.08	0.833**	0.109	.75			
Fo-Arbeiten	0.104**	0.021	.02	0.266**	0.033	.07	0.150**	0.023	.03
	R^2 = .87			R^2 = .66			R^2 = .84		

Legende. Koeffizient = Regressionskoeffizient; *SE* = Standardfehler. Validierungsvariablen: Lehramt = Dummy Studiengang Lehramt, Psychologie = Dummy Studiengang Psychologie (Referenzkategorie Studiengang Erziehungswissenschaft); Deutschnote: 1 = sehr gut, 2 = gut, 3 = befriedigend, 4 = ausreichend; e-papers = Häufigkeit des Lesens von e-papers (1 = sehr selten oder nie, 2 = einmal im Monat, 3 = einmal in der Woche, 4 = mehrmals in der Woche, 5 = täglich, 6 = mehrmals täglich); Fo-Artikel = verfasster Forschungsartikel (nein = 1, ja = 2); Semester 1–4 = Dummy Studierphase Anfänger, Promovierende = Dummy Studierphase Promovierende (Referenzkategorie Semester 5 und höher bzw. Master); Fo-Arbeiten = verfasste Forschungsarbeiten (maximal 5 Punkte, wenn folgende Arbeiten genannt wurden: Hausarbeit, Bachelorarbeit, Masterarbeit, wissenschaftliches Poster, sonstige Forschungsarbeiten).

Tab. 4: Latente Regression des Genrewissens auf unterschiedliche Lerngelegenheiten, die Studierphase und den Studiengang

nisse der Regression in den Modellen 2 und 3 jeweils unter Ausschluss einer der beiden Variablen präsentiert (Tab. 4).

In Tabelle 4 werden sowohl die unstandardisierten als auch die standardisierten Koeffizienten der latenten Regressionsanalyse berichtet. Die standardisierten Koeffizienten errechnen sich jeweils aus dem Produkt des Koeffizienten der Validierungsvariablen mit dem Quotienten des jeweiligen Standardfehlers der Variablen und dem Standardfehler der Modellkonstanten (Gelman, 2008). Die Standardisierung ermöglicht damit einen Vergleich der Regressionsgewichte der Variablen untereinander.

Modell 1, in das alle Variablen eingingen ($R^2 = .87$), und Modell 3, in dem die Variable der Studierphase ausgeschlossen war ($R^2 = .84$), klärten den höchsten Anteil an Varianz auf. Der dominante Einfluss des Studiengangs trat in diesen beiden Modellen deutlich hervor. In Modell 1 war der Studiengang der Psychologie, der zwei Drittel aller teilnehmenden Promovierenden umfasste, um mehr als eine Standardabweichung besser als der Lehramtsstudiengang, der wiederum fast eine halbe Standardabweichung schlechter abschnitt als der Referenzstudiengang Erziehungswissenschaft. Unter Ausschluss der Studierphase im Modell 3 zeigte sich der Einfluss des Studiengangs noch etwas deutlicher als in Modell 1. Ein geringerer Einfluss wurde bei der Variable Forschungsartikel über alle Modelle sichtbar. Im Rahmen von Modell 2 trat unter Ausschluss des Studiengangs die Studierphase hervor: Promovierende erreichten ein um fast eine Standardabweichung besseres Testergebnis als Studienanfängerinnen und Studienanfänger, deren Leistung nahezu vergleichbar mit derjenigen der Referenzkategorie der Semester 5 bis Master war. Neben der Variablen Studiengang mit dem deutlich stärksten Vorhersagewert übten demnach die Studierphase und das (Mit-)Verfassen von Forschungsartikeln einen weiteren Einfluss auf die abhängige Variable Genrewissen aus.

5. Diskussion

Unsere Ergebnisse zeigen, dass es uns gelungen scheint, ein reliables und testökonomisches Instrument zur Erfassung des Genrewissens *empirischer Forschungsartikel* zu entwickeln. Sowohl die Gütekriterien der Itemfits als auch die Experteneinschätzungen bezüglich deren Inhaltsvalidität legen nahe, dass die Items valide sind. Im Rahmen der Prüfung der Konstruktvalidität konnte gezeigt werden, dass deklaratives und diagnostisches Genrewissen über empirische Forschungsartikel in den untersuchten Studiengängen in einem unterschiedlichen Ausmaß vermittelt werden. Anhand dreier Modelle der latenten Regressionsanalyse wird deutlich, dass unser Test deutlich zwischen einzelnen Studierphasen und den Studiengängen Psychologie, Erziehungswissenschaft und Lehramt differenzieren kann, auch wenn Studierphase und Studiengang in unserer Stichprobe in beträchtlichem Maße korreliert waren.

Während die Zusammensetzung unserer Stichprobe in Hinblick auf die erfassten Studierenden auf den ersten Blick heterogen erscheint, ist die Berücksichtigung von Studierenden der Lehrämter durchaus begründet: Das Lehramtsstudium ist zwar für sich betrachtet kein genuin bildungswissenschaftlicher Studiengang, allerdings rekrutieren

sich aus dem Lehramt spätere Fachdidaktiker und Fachdidaktikerinnen. Außerdem weist das Lehramtsstudium erhebliche bildungswissenschaftliche Anteile auf, deren Umfang und Bedeutung im Zuge der Bemühungen um eine Reform der Lehrerbildung (siehe die KMK-Standards für Lehrerbildung, 2004) sowie der Diskussionen um eine stärkere Forschungsorientierung im Lehramtsstudium in den letzten Jahren enorm gewachsen ist (Borg, 2010; Niemi & Nevgi, 2014).

Bezüglich der Itemkennwerte könnten die Misfit-Werte im Bereich *diagnostisches rhetorisches Wissen über wissenschaftlichen Schreibstil* möglicherweise auf Items hinweisen, die ungeeignet sind, die latente Dimension Genrewissen valide zu erfassen. Diese Vermutung steht in Einklang mit gleichzeitig erhobenen Expertenurteilen, die deren Inhaltsvalidität uneinheitlich einschätzten. Diese Items lassen sich überwiegend unter die Wissensanforderungen zum wissenschaftlichen Schreibstil (Tab. 1) subsumieren, wie das Beispiel einer Satzauswahlaufgabe zeigt: Dabei sollten jeweils zwei Sätze ausgeschlossen werden, die mit unnötigen Substantivierungen, umständlicher passiver Satzkonstruktion, unnötigem Gebrauch von Fremdwörtern, grammatikalischen Fehlern, Wertungen sowie Füllwörtern angereichert worden waren. In Hinblick auf die Inhaltsvalidität sehen wir weitere Optimierungsmöglichkeiten bei einzelnen Items. Beispielsweise kommt es bei der Frage nach übergeordneten Bestandteilen eines empirischen Forschungsartikels während der Itemkodierung bei dem Begriff *Einleitung* zu dem Problem, dass sich ein alltagssprachlich gebräuchlicher Begriff ungünstig mit dessen fachwissenschaftlicher Bedeutung überschneidet. Somit könnte es speziell bei diesem Item sein, dass das Wissen von Studierenden tendenziell überschätzt wird. Wie zudem den Itemschwierigkeiten der klassischen Testtheorie zu entnehmen ist – bis auf zwei leichte Items liegen diese überwiegend im mittleren Schwierigkeitsbereich –, könnte die Nachentwicklung von Items sowohl im oberen (= schwierigen) als auch im unteren (= einfachen) Bereich das Testverfahren weiter verbessern.

Die hier vorgestellten 24 Items bilden dennoch eine erste gute Grundlage für unseren standardisierten, raschskalierten Test zu Genrewissen. Nach einer testökonomischen Diagnose (< 45 min) anhand unseres Instruments könnten somit gezielt Maßnahmen zur Förderung von Genrewissen und damit der Kompetenz zum wissenschaftlichen Schreiben erfolgen. Als geeigneter Einsatz wären studiengangübergreifende Evaluationen von Lernzielen in diesem Bereich denkbar. Beispielsweise legen die vorliegenden Ergebnisse nahe, dass Lehramtsstudierende bezüglich des expliziten Genrewissens einen Rückstand gegenüber Studierenden der Erziehungswissenschaft und der Psychologie aufweisen, obwohl in Hinblick auf die Lehrerausbildung auch von Lehramtsstudierenden die Fähigkeit zur Rezeption empirischer Forschungsartikel gefordert wird, eine Kompetenz also, die eine solide Basis an Genrewissen voraussetzt.

Bislang noch nicht erfassbar im Rahmen des Tests sind Anforderungen zur Anwendung rhetorischen Wissens beim Verfassen bzw. Produzieren von Artikelmanuskripten. Ungeachtet der ermutigenden Ergebnisse ist somit einschränkend festzuhalten, dass fundiertes theoretisches Genrewissen nicht automatisch eine gute Schreibfähigkeit und somit gelungene Artikelmanuskripte nach sich zieht. Deshalb untersuchen wir in einer laufenden Studie genauer die Rolle des Genrewissens bei der Textproduktion (im Zu-

sammenhang mit anderen Variablen wie etwa Schreibüberzeugungen und allgemeiner sozialwissenschaftlicher Forschungskompetenz). Wir versuchen dabei, unterschiedliche Qualitäten wissenschaftlicher Abstracts bzw. unterschiedliche Kompetenzniveaus des wissenschaftlichen Schreibens messbar zu machen und den Beitrag des Genrewissens, gemessen mit unserem hier vorgestellten Test, zur Vorhersage wissenschaftlicher Schreibkompetenz zu bestimmen.

Literatur

Adams, R. J. (2002). Scaling PISA cognitive data. In R. J. Adams & M. L. Wu (Hrsg.), *PISA 2000 technical report* (S. 99–108). Paris: OECD.

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher cognition and its relation to visual attention. *Journal Human-Computer Interaction, 12*(4), 439–462.

APA American Psychological Association (2009). *Publication manual of the American Psychological Association* (6. Aufl.). Washington, D. C.: American Psychological Association.

Backhaus, K., Erichson, B., & Weiber, R. (2011). *Fortgeschrittene multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. Heidelberg: Springer.

Bazerman, C. (1983). Scientific Writing as a Social Act. New Essays in Technical Writing and Communication. In P. V. Anderson, R. J. Brockman & C. R. Miller (Hrsg.), *New essays in technical and scientific communication* (S. 84–156). Farmingdale: Baywood,

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna & H. L. Roediger III. (Hrsg.), *The complete academic: A career guide, 2* (S. 185–219). Washington, D. C.: American Psychological Association.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. New York: Routledge.

Borg, S. (2010). Language teacher research engagement. *Language Teaching, 43*(4), 391–429.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.

DGP Deutsche Gesellschaft für Psychologie (2007). *Richtlinien zur Manuskriptgestaltung, 3*. Göttingen: Hogrefe.

Dittmann, J., Geneuss, K. A., Nennstiel, C., & Quast, N. A. (2003). Schreibprobleme im Studium – Eine empirische Untersuchung. In K. Ehlich & A. Steets (Hrsg.), *Wissenschaftlich schreiben – lehren und lernen* (S. 155–185). Berlin: de Gruyter.

Gelman, A. (2008). Scaling regression inputs by deviding by two standard deviations. *Statistics in Medicine, 27*, 2865–2873.

Goldman, S. R., & Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J. A. León & A. C. Graesser (Hrsg.), *The psychology of science text comprehension* (S. 417–436). Mahwah: Erlbaum.

Gruber, H., Muntigl, P., Reisigl, M., Rheindorf, M., Wetschanow, K., & Christine, C. (2006). *Genre, Habitus und wissenschaftliches Schreiben*. Münster: LIT.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Hrsg.), *Cognitive processes in writing: An interdisciplinary approach* (S. 3–30). Hillsdale: Lawrence Erlbaum Associates.

Jude, N. (2006). *IRT-Skalierung mit ConQuest*. http://media.metrik.de/uploads/incoming/pub/Literatur/Folien_Jude+anleitung%20zu%20ConQuest.pdf [23. 02. 2006].

Kamler, B., & Thomson, P. (2006). *Helping doctoral students write: Pedagogies for supervision*. London: Routledge.

KMK Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004). *Standards für die Lehrerbildung*. Bericht der Arbeitsgruppe, Ms 2004.

Kruse, O. (1997). Wissenschaftliche Textproduktion und Schreibdidaktik. In E.-M. Jakobs & D. Knorr (Hrsg.), *Schreiben in den Wissenschaften* (S. 141–158). Frankfurt a. M.: Lang.

Kruse, O., Jakobs, E.-M., & Ruhmann, G. (1999). *Schlüsselkompetenz Schreiben. Konzepte, Methoden, Projekte für Schreibberatung und Schreibdidaktik an der Hochschule* (S. 19–34). Neuwied/Kriftel: Luchterhand.

Muthén, L. K., & Muthén, B. O. (2001). *Mplus: The comprehensive modeling program for applied researchers user's guide*. Los Angeles: Muthén & Muthén.

Niemi, H., & Nevgi, A. (2014). Research studies and active learning promoting professional competences in Finnish teacher education. *Teaching and Teacher Education, 43*, 131–142.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.

Rijlaarsdam, G., Couzijn, M., Janssen, T., Braaksma, M., & Kieft, M. (2006). Writing Experiment Manuals in Science Education: The impact of writing, genre, and audience. *International Journal of Science Education, 28*(2-3), 203–233.

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

Schindler, K. (2008). *Wissenschaftliches Schreiben in Sprach- und Kommunikationswissenschaft – Zwei Beispiele für schreibintensive Lehrveranstaltungen in den Geisteswissenschaften*. Köln: Zeitschrift Schreiben.

Silvia, P. J. (2007). *How to write a lot: A practical guide to productive academic writing*. Washington, D. C.: American Psychological Association.

Simonton, D. K. (2003). Expertise, Competence and Creative Ability: The Perplexing Complexities. In R. J. Sternberg & E. L. Grigorenko (Hrsg.), *The Psychologiy of Abilities, Competencies, and Expertise* (S. 213–239). Cambridge: Cambridge University Press.

Strunk, W. Jr., & White, E. B. (1999). *The Elements of Style, 4*. New York: Pearson Longman.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. New York: Cambridge University Press.

Wilson, M. R. (2005). *Constructing Measures: An item response modeling approach*. Mahwah: Lawrence Erlbaum.

Wirtz, M. A., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe/Verlag für Psychologie.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest: Version 2.0. Generalised Item Response Modelling Software*. Camberwell: Acer.

Abstract: In the Educational Sciences, knowledge about the genre *empirical research article* is a core component of the competence to write scientific texts. As diagnostic tools for assessing students' rhetorical knowledge and skills are lacking, our aim was to develop such a test for the Educational Sciences. Undergraduate and graduate students (N = 372) from Educational Science, Psychology and teacher education programs answered 37 questions about the genre *empirical research article*. Based on a dichotomous Rasch model (IRT), all items were scaled and fitting items were identified. 24 Items met our criteria. This selection of items represents an economic test that reliably distinguishes between different study stages and programs. The test could be used to gain insight into how study programs differ regarding the successful instruction of rhetorical knowledge and skills in the domain of scientific writing.

Keywords: Genre Knowledge, Competency, Scientific Writing, Assessment, IRT

Anschrift der Autor(inn)en

Alexandra Winter-Hölzl, Albert-Ludwigs-Universität Freiburg, Institut für Erziehungswissenschaft, Rempartstraße 11, 79085 Freiburg, Deutschland
E-Mail: alexandra.winter@ezw.uni-freiburg.de

Dr. Kristin Wäschle, Albert-Ludwigs-Universität Freiburg, Institut für Erziehungswissenschaft, Rempartstraße 11, 79085 Freiburg, Deutschland
E-Mail: kristin.waeschle@ezw.uni-freiburg.de

Prof. Dr. Jörg Wittwer, Albert-Ludwigs-Universität Freiburg, Institut für Erziehungswissenschaft, Rempartstraße 11, 79085 Freiburg, Deutschland
E-Mail: joerg.wittwer@ezw.uni-freiburg.de

Prof. Dr. Rainer Watermann, Freie Universität Berlin, Empirische Bildungsforschung, Habelschwerdter Allee 45, 14195 Berlin, Deutschland
E-Mail: rainer.watermann@fu-berlin.de

Prof. Dr. Matthias Nückles, Albert-Ludwigs-Universität Freiburg, Institut für Erziehungswissenschaft, Rempartstraße 11, 79085 Freiburg, Deutschland
E-Mail: matthias.nueckles@ezw.uni-freiburg.de

Gabriele Steuer/Tobias Engelschalk/Gregor Jöstl/Anne Roth/ Bastian Wimmer/Bernhard Schmitz/Barbara Schober/Christiane Spiel/Albert Ziegler/Markus Dresel

Kompetenzen zum selbstregulierten Lernen im Studium

Ergebnisse der Befragung von Expert(inn)en aus vier Studienbereichen[1]

Zusammenfassung: Kompetenzen zum selbstregulierten Lernen (SRL) gelten als zentrale Voraussetzung des Studienerfolgs. Bisher ist unklar, in welchen Situationen der Einsatz welcher SRL-Strategien als kompetent anzusehen ist. Auf Basis einer Onlinebefragung von 306 Expert(inn)en für das SRL werden im Beitrag quantitative Evidenzen zur relativen Bedeutung sowie zu den selbstregulativen Anforderungen unterschiedlicher Lernsituationen in vier Studienbereichen dargestellt. Ferner werden Expert(inn)enurteile zur relativen Eignung von SRL-Strategien in diesen Situationen berichtet. Die untersuchten Strategien wurden aus einem differenziert und umfassend konzipierten Arbeitsmodell abgeleitet, das einen integrativen Ansatz zur Konzeption und Struktur von SRL-Kompetenzen im tertiären Bereich liefert.

Schlagworte: selbstreguliertes Lernen, Kompetenzerfassung, Lernsituationen, Lernstrategien, Expert(inn)enbefragung

1. Theoretischer Hintergrund

An der Hochschule wird erwartet, dass sich Studierende – unabhängig davon, welches Fach sie studieren – umfangreiche Wissensbestände eigenständig erarbeiten. Daher gelten Kompetenzen zum selbstregulierten Lernen (SRL) als generische, d. h. bereichsübergreifende Kompetenzen, die im Studium von hoher Bedeutung sind. Dazu zählen beispielsweise Kompetenzen zur effektiven Aneignung neuer Inhalte, zur Überwachung des eigenen Lernfortschritts oder zur Überwindung von Motivationsproblemen beim Lernen. Die Förderung dieser Kompetenzen wird auch explizit als wichtiges Ziel tertiärer Bildung betrachtet, was etwa im Qualifikationsrahmen für deutsche Hochschulabschlüsse zum Ausdruck kommt (Ständige Kultusministerkonferenz der Länder, 2005). Kompetenzen zum selbstregulierten Lernen können damit als eine Voraussetzung, aber auch als ein Ergebnis eines erfolgreichen Studiums betrachtet werden.

1 Die in dieser Arbeit präsentierte Forschung wurde durch Mittel des Bundesministeriums für Bildung und Forschung an Markus Dresel (01 PK11020A), Bernhard Schmitz (01 PK11020B), Barbara Schober und Christine Spiel (01 PK11020C) sowie Albert Ziegler (01 PK11020D) gefördert.

Die bisherige Forschung bezieht sich überwiegend auf SRL auf Verhaltensebene. Unter idealem selbstregulierten Lernen wird dabei meist verstanden, dass Lernende ihr eigenes Lernen autonom und sachangemessen steuern, d. h. sich eigenständig passende Ziele für ihr Lernen setzen, Lernstrategien anwenden, die zu Zielen und Lerngegenständen passen, ihre Motivation auch bei Widrigkeiten aufrechterhalten, ihr Lernen – wenn notwendig – anpassen und Zielerreichung wie Lernergebnisse während und nach Abschluss des Lernprozesses bewerten (vgl. Artelt, Demmrich & Baumert, 2001). Mittlerweile liegen umfangreiche Forschungsergebnisse zu selbstreguliertem Lernen auf behavioraler Ebene vor, die die Bedeutung der genannten Regulationsprozesse beim Lernen unterstreichen (vgl. Zimmerman & Schunk, 2011).

Im Gegensatz zu der breiten SRL-Forschung auf Verhaltensebene mangelt es jedoch noch weitgehend an theoretisch fundierten und empirisch abgesicherten Ansätzen zur Beschreibung und validen Erfassung der Kompetenzen, die zur effektiven Selbstregulation erforderlich sind – dies gilt insbesondere für SRL im Hochschulbereich, an das aufgrund komplexer Lernaufgaben bei gleichzeitig großer Autonomie größere Anforderungen gestellt werden als etwa im schulischen Bereich (vgl. Händel, Artelt & Weinert, 2013; Schlagmüller & Schneider, 2007; Wirth & Leutner, 2008). Konzeptuell-theoretisch gehen Kompetenzen zum SRL dabei deutlich über die reine Realisierung von SRL auf Verhaltensebene, z. B. die Anwendung von Lernstrategien, hinaus. Sie stellen die dahinterstehenden personalen Dispositionen für das Regulationsverhalten dar, die erforderlich sind, um gelingendes SRL auf Verhaltensebene zeigen zu können (= Performanz).

Entsprechend der theoretischen Unklarheiten bei der Konzeption von SRL-Kompetenzen von Studierenden bestehen auch größere Forschungsdefizite in Bezug auf deren Messung. Ziel der vorliegenden Arbeit ist es deshalb, Grundlagen für die zukünftige Entwicklung von Messinstrumenten zur Erfassung der SRL-Kompetenzen von Studierenden zu legen, die insbesondere zu deren Konzeption und zur Sicherung ihrer Inhaltsvalidität genutzt werden können (vgl. Jenßen, Dunekacke & Blömeke, 2015, in diesem Beiheft). Diese betreffen im Einzelnen die Lernsituationen in unterschiedlichen Studienfächern, in denen Kompetenzen des SRL in besonderem Maße erforderlich sind, und die relative Eignung verschiedener SRL-Strategien in diesen Situationen. Evidenzen dazu wurden mithilfe der quantitativen Befragung von 306 Praxis-Expert(inn)en für SRL im Studium (Dozierende und exzellente Studierende) an vier verschiedenen Hochschulstandorten in vier unterschiedlichen Studienbereichen (Wirtschaftswissenschaften, E-Technik, Lehrer(innen)bildung MINT, Psychologie) gewonnen. Diese Evidenzen können nicht nur im Zusammenhang der Messinstrumenteentwicklung genutzt werden, sondern auch zur Weiterentwicklung des theoretischen Verständnisses des SRL und, beispielsweise, daraus abgeleiteter Fördermaßnahmen.

1.1 Integratives Strukturmodell der Kompetenzen zum SRL

Als Basis der vorliegenden Arbeit entwickelten wir ein Arbeitsmodell der SRL-Kompetenzen von Studierenden, das etablierte theoretische Perspektiven der SRL-Forschung verbindet sowie bewusst umfassend und differenziert im Hinblick auf mögliche Kompetenzfacetten ist (s. a. Dresel et al., im Druck; Schober et al., im Druck). Dieses integrative Strukturmodell (Abb. 1) wird durch die Kombination von drei Dimensionen mit jeweils drei Differenzierungen aufgespannt: die Strategiedimension (kognitive Lernstrategien, metakognitive Strategien und ressourcenbezogene Strategien), die Wissensdimension (deklaratives, prozedurales und konditionales Strategiewissen) und die Prozessdimension (präaktionale, aktionale und postaktionale Phase).

Die drei Dimensionen leiten sich aus bewährten theoretischen SRL-Modellen ab, die eine systematische Konzeptualisierung von Komponenten (Bestandteilen von SRL) und Prozessen (Abläufen beim SRL) anstreben (z. B. Boekaerts, Pintrich & Zeidner, 2000; Schunk & Zimmerman, 2003; Ziegler, Porath & Stöger, 2011). Die vorliegenden SRL-Modelle lassen sich daher zwei Gruppen zuordnen: komponentenorientierte und prozessorientierte Modelle (vgl. Thillmann, 2007; Winne & Perry, 2000). Beide Perspektiven stellen für die Modellierung und Messung von Kompetenzen zum SRL im tertiären Bereich eine wichtige Grundlage dar und wurden deshalb im Arbeitsmodell integriert.

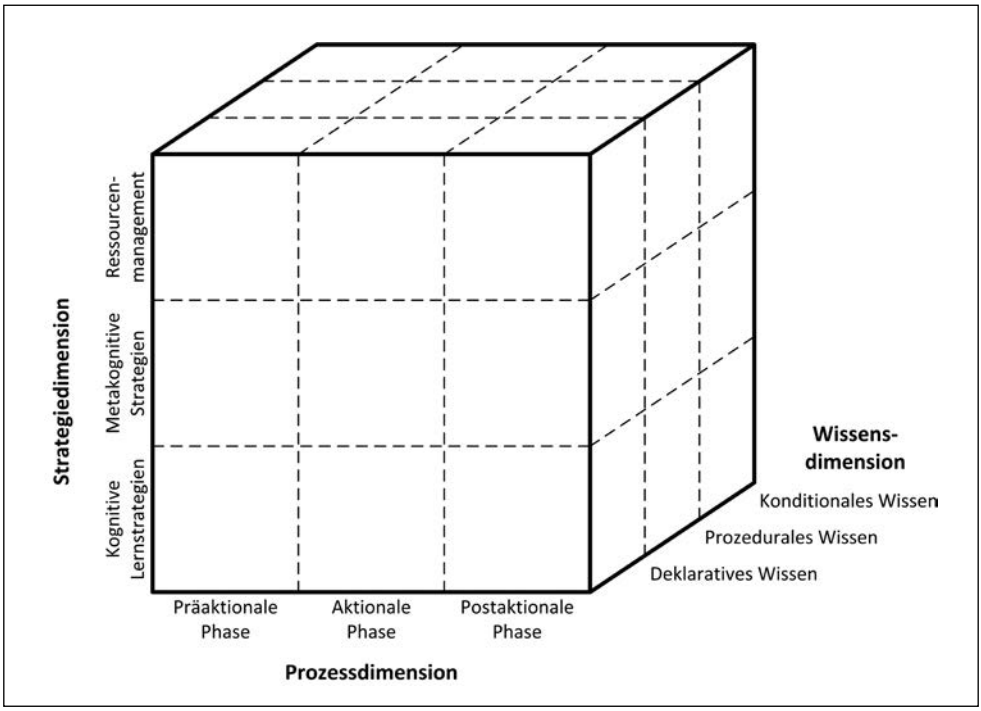


Abb. 1: Strukturmodell der Kompetenzen zum selbstregulierten Lernen

Komponentenorientierte Modelle erheben den Anspruch, alle am SRL beteiligten Komponenten zu identifizieren. Hier kann einerseits die Art der SRL-Strategie selbst kategorisiert werden (Strategiedimension) und andererseits die Form, in der das Strategiewissen vorliegt (Wissensdimension). Entlang der Strategiedimension werden die SRL-Strategien nach ihrer Art in drei grobe Kategorien unterschieden: kognitive Lernstrategien, metakognitive Strategien und ressourcenorientierte Strategien (vgl. Boekaerts & Corno, 2005). Diese drei groben Kategorien lassen sich jeweils nochmals feiner untergliedern (ist in Abb. 1 nicht explizit dargestellt). So werden kognitive Strategien meist nochmals anhand ihrer anvisierten Funktionen (häufig: Memorieren, Elaborieren, Organisieren; z. B. Weinstein & Hume, 1998) und/oder ihrer Verarbeitungstiefe (oberflächenorientierte vs. tiefenorientierte Strategien; z. B. Marton & Saljö, 1984) unterteilt. Metakognitive Strategien werden häufig in Zielsetzung und Planung, Selbstüberwachung bzw. Kontrolle sowie Regulation und ggf. Reflexion differenziert (Boekaerts, 1999; Friedrich & Mandl, 1995; Wild & Schiefele, 1994). Ressourcenbezogene Strategien umfassen die Regulation interner (z. B. Motivations- und Emotionsregulation) und externer (z. B. Nutzung zusätzlicher Informationsquellen, Hilfesuche) Ressourcen (vgl. Wild & Schiefele, 1994). Bezüglich der kognitiven und auch der metakognitiven Strategien liegt eine Vielzahl von Forschungsarbeiten vor, die ihre Relevanz im Lernprozess unterstreichen, wohingegen bei den ressourcenbezogenen Strategien noch erhebliche Defizite (Überblick bei Zimmerman & Schunk, 2011) bestehen. Bisherige Befunde zur Motivations- und Emotionsregulation weisen jedoch auf ihre Bedeutsamkeit hin (z. B. Leutner, Barthel & Schreiber, 2001; Pekrun, Goetz, Titz & Perry, 2002; Schwinger, Steinmayr & Spinath, 2009; Wolters, 2003).

Es ist anzunehmen, dass sich Kompetenzen von Lernenden zum SRL nicht nur auf der Grundlage unterschiedlicher Arten von SRL-Strategien konstituieren, sondern auch auf der Basis des Umfangs, der Anwendbarkeit und der Angemessenheit des diesbezüglichen Strategiewissens (Wissensdimension). So können SRL-Strategien in unterschiedlichen Wissensformen vorliegen: als deklaratives, prozedurales oder konditionales Strategiewissen (vgl. Paris, Lipson & Wixson, 1983; Shraw & Moshman, 1995). Bei deklarativem Strategiewissen handelt es sich um Faktenwissen über Strategien (z. B. Kenntnis von Lernstrategien), prozedurales Strategiewissen umfasst Handlungswissen zum Strategieeinsatz, und konditionales Strategiewissen bezieht sich schließlich auf Wissen über die Passung von SRL-Strategien und Lern-/Regulationsanforderungen, d. h. über die Eignung verschiedener SRL-Strategien in verschiedenen Situationen. Evidenzen dafür, dass Strategiewissen in den drei beschriebenen Wissensarten vorliegt, ergaben sich beispielsweise in der Forschung zur Entwicklung von Lernstrategien im Kindes- und Jugendalter (z. B. Lehmann & Hasselhorn, 2009; Schneider & Lockl, 2006), aber auch zum SRL in der Sekundarstufe (z. B. Artelt, 2000; Dresel & Haugwitz, 2005). Bisherige Untersuchungen zeigten beispielsweise, dass zum Teil gravierende Diskrepanzen zwischen dem (oft umfangreichen) deklarativen Wissen und der (oft defizitären) tatsächlichen Anwendung von SRL-Strategien bestehen (vgl. Artelt, 2000). Ein Beispiel für die Kombination von Strategie- und Wissensdimension wäre, dass bei der Planung von Lernaktivitäten (metakognitive Strategien) Wissen darüber vorhanden sein muss,

welche kognitiven Lernstrategien unter welchen Bedingungen effektiv sind (konditionales Strategiewissen).

In prozessorientierten Modellen des SRL werden Lernprozesse in dynamischen Modellen mit rekursiven Lernzyklen beschrieben (z. B. Schiefele & Pekrun, 1996; Schmitz & Wiese, 2006; Winne & Hadwin, 1998; Ziegler, Stöger & Dresel, 2004; Zimmerman, 2000). Dabei werden meist mindestens drei Phasen im Lernprozess unterschieden (vgl. Heckhausen & Gollwitzer, 1987; Schmitz & Wiese, 2006; Zimmerman, 2000): die präaktionale Phase, die aktionale Phase und die postaktionale Phase (Prozessdimension). Für die jeweiligen Phasen werden in den entsprechenden Modellen Anforderungen an Lernende formuliert, nicht jedoch die Kompetenzen dazu benannt, wie diesen begegnet werden kann (vgl. Wirth & Leutner, 2008). Dabei ist davon auszugehen, dass in jeder der drei Prozessphasen die Kompetenzaspekte von Bedeutung sind, die mit der Strategie- und der Wissensdimension benannt wurden. Dies bedeutet, dass in den drei Phasen unterschiedliche Strategiearten angewendet werden können und sollten, wobei die Qualität dieser Strategieanwendung auch von der Form ihrer wissensmäßigen Repräsentation abhängig ist.

Eine Grundannahme, die der Formulierung unseres Arbeitsmodells zugrunde lag, ist, dass sich die Kompetenzen von Studierenden zum SRL anhand der Kombination der drei dargestellten Dimensionen des SRL hinreichend umfassend beschreiben lassen (vgl. Abb. 1). Die simultane Berücksichtigung dieser drei Dimensionen, die in der bisherigen Forschung kaum vereinigt wurden, ermöglicht es aus unserer Sicht, über die rein behaviorale Ebene des Strategieeinsatzes hinauszugehen. Aus der Kombination der je drei groben Ausprägungen der drei Dimensionen ergäben sich prinzipiell 27 Zellen, die potenzielle Kompetenzfacetten repräsentieren. Allerdings ist anzumerken, dass Kompetenzen zu kognitiven Lernstrategien nur für die aktionale Phase, nicht aber für die präaktionale oder postaktionale Phase plausibel begründbar sind. Entsprechend ergeben sich 21 theoretisch sinnvolle Zellen im Arbeitsmodell, die potenziellen Kompetenzfacetten entsprechen.

1.2 Erfassung von SRL-Kompetenzen

Ein stärkerer theoretischer Fokus auf die erfolgreichem SRL zugrunde liegenden Kompetenzen hat das Potenzial, die SRL-Forschung insgesamt zu befruchten (vgl. Wirth & Leutner, 2008). Dazu sind valide Messinstrumente zur Erfassung von SRL-Kompetenzen erforderlich. Der Anspruch, Kompetenzen zum SRL zu erfassen, stellt dabei weit höhere methodische Anforderungen (vgl. Jenßen et al., 2015, in diesem Beiheft) als die Messung von SRL auf Verhaltensebene. Gleichzeitig besteht aber die begründete Hoffnung, damit Limitationen bisheriger Erfassungsinstrumente (insb. unspezifischer Selbstberichtsverfahren, die oft nur in geringem Maße mit Außenkriterien wie Leistung im Zusammenhang stehen) zu überwinden (für Überblicke siehe Spörer & Brunstein, 2006; Wirth & Leutner, 2008). Die Hoffnung, mit der validen Erfassung von SRL-Kompetenzen die in der SRL-Forschung gut dokumentierte Validitätsproblematik abzumil-

dern, gründet insbesondere auf den bereits erwähnten Passungsgedanken: Theoretisch ist nicht davon auszugehen, dass ein Maximum an Einsatz aller SRL-Strategien in allen Situationen funktional für den Lernprozess ist („maximum view" sensu Wirth & Leutner, 2008), sondern dass in bestimmten Lernsituationen nur bestimmte Lernstrategien geeignet sind und andere nicht.

Im tertiären Bereich wird selbstreguliertes Lernen allerdings bisher überwiegend durch globale Fragebogeninstrumente erfasst (z. B. MSLQ: Pintrich, Smith, Garcia & McKeachie, 1991; LIST: Wild, Schiefele & Winteler, 1992). In solchen Instrumenten werden die unterschiedlichen Lernsituationen im Studium nicht benannt, sondern die befragten Personen müssen bei der Beantwortung implizit über verschiedene Situationen generalisieren. Dadurch sind die Messungen situationsunspezifisch – es ist davon auszugehen, dass manche Situationen vernachlässigt werden oder auch, dass die Messungen insofern nicht vergleichbar sind, da Studierende unterschiedliche Situationen bei der Beantwortung erinnern. Darüber hinaus wird bei globalen Selbstberichten häufig eher vom Grundsatz „viel hilft viel" (maximum view) ausgegangen als von einer situations- und anforderungsangemessenen Intensität des Strategieeinsatzes. Situationsunspezifische Selbstberichte sind folglich nicht dazu geeignet, die Passung des Strategieeinsatzes zu erfassen – dabei muss Wissen über die Passung von Strategien zu Anforderungssituationen als ein theoretischer Nukleus des Konzepts der SRL-Kompetenzen aufgefasst werden.

Wirth und Leutner (2008) sprechen in diesem Zusammenhang von qualitativen Standards, anhand derer die Passung des SRL zur spezifischen Anforderungssituation beurteilt werden kann. Solche qualitativen Standards können insbesondere aus Urteilen von Expert(inn)en abgeleitet werden. Mit der vorliegenden Arbeit sollen Grundlagen für die Definition von qualitativen Standards für die ökologisch valide Erfassung von SRL-Kompetenzen geschaffen werden.

1.3 Relevante und anforderungsreiche Lernsituationen in verschiedenen Studiengängen

Um Aussagen über die Passung von SRL-Strategien zu Anforderungssituationen zu generieren, ist zunächst belastbares Wissen darüber erforderlich, inwieweit verschiedene Lernsituationen in unterschiedlichen Studienfächern von Bedeutung für den Studienerfolg sind und – insbesondere – welche Anforderungen an SRL damit einhergehen. Es kann gefordert werden, dass eine ökologisch valide Erfassung von SRL-Kompetenzen im Studium die Passung von Strategien nicht zu beliebigen Lernsituationen, sondern zu relevanten und im Hinblick auf SRL anforderungsreichen Lernsituationen fokussieren sollte. Tatsächlich mangelt es jedoch bereits an einer empirisch fundierten Beschreibung entsprechender Lernsituationen im Studium. Dies wird dadurch verschärft, dass in unterschiedlichen Studienbereichen augenscheinlich ähnliche Lernsituationen unterschiedlich bedeutsam für den Studienerfolg sein können und unterschiedliche Erfordernisse der Selbstregulation des Lernens mit sich bringen können.

In einer der wenigen Studien zum Thema wurden österreichische Studierende zu wichtigen Situationen im Studium befragt (Zauchner, Baumgartner, Blaschitz & Weissenback, 2008). Ihre Antworten wurden mittels qualitativer Inhaltsanalyse ausgewertet. Vier Situationen wurden extrahiert: „Vorbereitung auf eine Prüfung", „praktische Erfahrungen sammeln", „etwas lesen" und „mit anderen diskutieren". Die beiden ersten Situationen wurden als formelle, die beiden übrigen als informelle Lernsituationen klassifiziert. Allerdings liegen keine empirischen Evidenzen dazu vor, welche Lernsituationen in unterschiedlichen Studienbereichen relevant für den Studienerfolg sind und gleichzeitig hohe Anforderungen an die Selbstregulation stellen. Dies ist jedoch eine notwendige Basis für die Bewertung und Interpretation der Befunde sowie für eine differenziertere Analyse von SRL-Kompetenzen. Eine empirische Untersuchung von typischen Lernsituationen im Studium ist umso wichtiger, als angenommen werden muss, dass hier Spezifika unterschiedlicher Studienfelder zum Tragen kommen können.

Erste, ebenfalls aber rein qualitative Hinweise dazu liefert eine eigene Vorarbeit (Dresel et al., im Druck). Darin wurden mittels halb-standardisierter Interviews insgesamt 108 Expert(inn)en (39 Dozierende und 69 als exzellent nominierte Studierende höheren Semesters) aus vier verschiedenen Studienbereichen nach den für den Studienerfolg bedeutsamen Lernsituationen und in diesen Situationen geeigneten SRL-Strategien offen befragt. Die qualitativen Interviews erbrachten Hinweise darauf, dass es über verschiedene Studienbereiche hinweg eine Reihe an strukturell ähnlichen Lernsituationen gibt, die sich drei Situationsgruppen zuordnen lassen: typische Selbstlernsituationen (z. B. Prüfungsvorbereitung), Erledigung selbstorganisierter, komplexer Anforderungen (z. B. Anfertigen einer wissenschaftlichen Arbeit) und Besuch sowie Vor- und Nachbereitung mehr oder weniger strukturierter Lerngelegenheiten (z. B. Vorlesungsbesuch). Auch wenn die Gemeinsamkeiten zwischen den Studienfächern eher überwogen als die Unterschiede, zeigten sich bei einigen Situationen kleine bis moderate Unterschiede in der Nennungshäufigkeit. Im Hinblick auf die in den spezifischen Situationen geeigneten SRL-Strategien erbrachte die Untersuchung Hinweise auf eine mittlere Situationsspezifität – u. a. wurden metakognitive Strategien und Strategien des Ressourcenmanagements in Situationen mit hohen SRL-Anforderungen als bedeutsamer bewertet. Des Weiteren wurden Strategien, die der präaktionalen oder der aktionalen Phase des Lernprozesses zugeordnet werden können, deutlich häufiger genannt als Strategien der postaktionalen Phase.

Diese Ergebnisse liefern qualitative Hinweise zur Relevanz unterschiedlicher Lernsituationen und zu den Anforderungen an SRL in unterschiedlichen Studienbereichen. Aufgrund des qualitativen Vorgehens lassen sie jedoch nur rudimentäre Schlussfolgerungen zur relativen Bedeutung der Lernsituationen und Strategien zu. Eine systematische Erfassung von Expert(inn)enurteilen mittels quantitativer Methoden sehen wir als sinnvolle Weiterführung an. Sie ist auch deshalb überfällig, da nach unserem Kenntnisstand bisher keine systematische Sammlung dazu vorliegt, welche Lernsituationen in unterschiedlichen Studienbereichen besonders anforderungsreich im Hinblick auf die Selbstregulation des Lernens sind (und sich deshalb besonders für eine situationsspezifische Erfassung von SRL anbieten).

1.4 Forschungsfragen

Übergeordnetes Ziel der in diesem Beitrag vorgestellten Befragung von Expert(inn)en für die Praxis des SRL im tertiären Bereich ist es, belastbare quantitative Evidenzen zur relativen Bedeutung unterschiedlicher Lernsituationen in unterschiedlichen Studienbereichen und zur relativen Eignung unterschiedlicher SRL-Strategien in diesen Situationen zu generieren – wobei sich die dabei adressierten SRL-Strategien im vorgeschlagenen Arbeitsmodell verorten lassen. Im Einzelnen sollen (1) Lernsituationen im Studium, zu deren erfolgreicher Bewältigung Kompetenzen zum selbstregulierten Lernen erforderlich sind, untersucht werden. Dabei soll auf die Bedeutsamkeit für ein erfolgreiches Studium sowie die Anforderungen an die Selbstregulation des Lernens in diesem spezifischen Set an Situationen abgehoben werden. Des Weiteren sollen (2) die Eignung bzw. Passung von SRL-Strategien in den einzelnen Situationen untersucht werden. Dabei erwarteten wir, dass sich passende Strategien über alle Ausprägungen von Strategie- und Prozessdimension verteilen. Die Wissensdimension war in der Untersuchung implizit eingebettet: Mit dem zentralen Fokus auf die Eignung bzw. Passung von SRL-Strategien wurden Evidenzen gewonnen, die primär zur Beurteilung von konditionalem Strategiewissen herangezogen werden können (vgl. Paris et al., 1983). Da geeignete Strategien offensichtlich auch gekannt und angemessen eingesetzt werden müssen, hat dies auch unmittelbare Konsequenzen für die Beurteilung von deklarativem und prozeduralem Strategiewissen. Als Querschnittsaspekt soll im Hinblick auf beide Forschungsfragen geprüft werden, ob und inwieweit Studienfachspezifika bestehen.

2. Methode

Zur Beantwortung der Forschungsfragen wurden quantitative Expert(inn)enbefragungen mittels standardisierter Online-Fragebögen durchgeführt.

2.1 Stichprobe

Um die Generalisierbarkeit der Befunde zu verbessern und sowohl das Lehren als auch das Lernen als relevante Praxisperspektiven auf SRL im tertiären Bereich berücksichtigen zu können, wurden als Expert(inn)en für die Praxis des SRL zwei Personengruppen definiert: Einerseits aktiv mit der Lehre befasste Wissenschaftler(innen) – dabei wurde davon ausgegangen, dass gerade Lehrende aufgrund ihrer Erfahrungen, wie Studierende mit unterschiedlichen Lernangeboten umgehen, über profundes Wissen zu relevanten Prozessen des SRL verfügen. Andererseits Studierende, die von den Dozierenden als exzellent nominiert wurden – dem lag die Annahme zugrunde, dass Kompetenzen zum SRL eine bedeutsame Rolle für den Studienerfolg spielen und dementsprechend in der Gruppe der besonders erfolgreichen Studierenden ein hoher Grad an SRL-Kompetenzen vorliegt. Einbezogen wurden Expert(inn)en aus vier Studienbereichen, nämlich der

Lehrer(innen)bildung in MINT-Fächern (Mathematik, Informatik, Naturwissenschaften, Technik), der Psychologie, den Wirtschaftswissenschaften und der Elektrotechnik. Um die Generalisierbarkeit der Ergebnisse noch weiter zu verbessern, wurde die Untersuchung zudem verteilt über vier Hochschulstandorte durchgeführt. Dabei wurden an jedem Standort zwei Studienbereiche berücksichtigt.

Die finale Expert(inn)enstichprobe umfasste insgesamt $N = 306$ Personen. Diese setzen sich zusammen aus $N = 144$ Dozierenden (Alter in Jahren: $M = 37.5$; $SD = 10.9$; Frauenanteil: 34.7 %; Lehrerfahrung an der Universität in Jahren: $M = 9.8$; $SD = 10.0$; derzeitiges Lehrdeputat in Semesterwochenstunden: $M = 5.6$; $SD = 4.5$) sowie $N = 162$ als exzellent nominierten Studierenden (Alter in Jahren: $M = 23.6$; $SD = 4.1$; Frauenanteil: 56.8 %; Abiturnote: $M = 1.9$; $SD = 0.6$; bisherige Noten im Studium: $M = 1.6$; $SD = 0.5$). Die Verteilung der Stichprobe über die beiden Expert(inn)engruppen, die vier Studienbereiche und die vier Hochschulstandorte ist in Tabelle 1 ersichtlich.

2.2 Vorgehen

Auf Basis der von Dresel et al. (im Druck) vorgelegten qualitativen Hinweise wurden Vignetten für insgesamt sieben typische Lernsituationen im Studium generiert (eine Auflistung findet sich in Tab. 2). Dazu wurden aus den qualitativen Daten – nach vorher festgelegten Regeln – Typikalitätsmerkmale der jeweiligen Situationen herausgearbeitet und kurze, möglichst prägnante Situationsbeschreibungen zusammengestellt. Die Satzstruktur wurde vereinheitlicht, um eine Äquivalenz der Expert(inn)enurteile zu gewährleisten. Im Folgenden ein Beispiel für die Situation „Prüfungsvorbereitung“: „Denken Sie an eine Vorbereitung für eine Prüfung. Damit ist eine Situation gemeint, in der große Stoffmengen selbständig aufbereitet und gelernt werden müssen. Dabei ist das Verständnis des Lernstoffs wesentlich. Die Prüfungsvorbereitung kann auch in der Lerngruppe stattfinden“. In der Onlinebefragung wurden den Expert(inn)en die Situa-

	Universität A			**Universität B**			**Universität C**			**Universität D**			**Gesamt**		
	D	**S**	**G**	**D**	**S**	**G**	**D**	**S**	**G**	**D**	**S**	**G**	**D**	**S**	**G**
Lehrer(innen)bildung in MINT-Fächern	18	20	38	–	–	–	12	20	32	–	–	–	30	40	70
Psychologie	–	–	–	17	20	37	–	–	–	24	21	45	41	41	82
Wirtschaftswissenschaften	28	21	49	–	–	–	–	–	–	21	20	41	49	41	90
Elektrotechnik	–	–	–	12	20	32	12	20	32	–	–	–	24	40	64
Gesamt	46	41	87	29	40	69	24	40	64	45	41	86	144	162	306

Anmerkungen. D = Dozierende, S = exzellente Studierende, G = gesamt.

Tab. 1: Überblick über die Stichprobe

tionsvignetten in randomisierter Reihenfolge vorgelegt. Neben der Relevanz für den Studienerfolg (mit dem Item „Wie bedeutsam ist diese Situation für ein erfolgreiches Studium in Ihrem Fach?“) sollten die Expert(inn)en auch den Anforderungsgehalt der jeweiligen Situation in Bezug auf SRL („Wie hoch ist in der oben beschriebenen Situation die Anforderung für Studierende Ihres Fachs, ihr Lernen selbst zu steuern/regulieren?“) einschätzen. Die Urteile wurden jeweils auf Likert-Skalen erfasst, die von 1 *(gar nicht bedeutsam* bzw. *sehr gering)* bis 6 *(sehr bedeutsam* bzw. *sehr hoch)* reichten.

In einem zweiten Schritt wurden die Expert(inn)en gebeten, insgesamt neun Gruppen von Selbstregulationsstrategien im Hinblick auf ihre Bedeutsamkeit für die Bewältigung von jeder der präsentierten sieben Lernsituationen einzuschätzen („Wie wichtig ist es, die folgenden Strategien einzusetzen, um in der oben genannten Situation erfolgreich zu sein?“). Die Items wurden dabei deduktiv aus den Taxonomien für SRL-Strategien (Komponenten- und Prozessmodelle, d. h. Strategie- und Prozessdimension im Arbeitsmodell) abgeleitet und auf spezifischer Ebene durch markante Strategien der interviewbasierten Vorarbeit ergänzt (vgl. Dresel et al., im Druck). Die Strategiebeschreibungen – die immer aus einem Schlagwort und zwei bis drei typischen Beispielen bestanden – wurden auf einem mittleren Abstraktionsniveau der Strategien formuliert und wiederum formal-grammatikalisch parallelisiert, um ihre Äquivalenz zu gewährleisten. Zwei der Beschreibungen bezogen sich auf kognitive Strategien (aktionale Phase), vier auf metakognitive Strategien (Planung in der präaktionalen, Überwachung und Regulation in der aktionalen sowie Reflexion in der postaktionalen Phase) und die verbleibenden drei auf ressourcenbezogene Strategien (Motivations- und Emotionsregulation gleichermaßen in allen drei Phasen, Strategien zur Regulation externer Ressourcen in der präaktionalen und aktionalen Phase). Eine Beispielbeschreibung lautet (Gruppe der Überwachungsstrategien als Teilgruppe der metakognitiven Strategien in der aktionalen Phase): „Strategien zur Überwachung des Lernens (z. B. Überwachung des Lernfortschritts, Überwachung des Einsatzes von Strategien)“. Die insgesamt 9 (Strategiegruppen) × 7 (Lernsituationen) Expert(inn)enurteile wurden auf Likert-Skalen erfasst, die von 1 *(gar nicht wichtig)* bis 10 *(sehr wichtig)* reichten. Im Anschluss an die Einzelstrategieurteile sollten zusätzlich die drei aus Experten(inn)ensicht wichtigsten SRL-Strategien in der jeweiligen Situation identifiziert werden. Diese Prioritätswahlen dienten der zusätzlichen Überprüfung der relativen Wichtigkeit der Strategien.

3. Ergebnisse

3.1 Typische Lernsituationen im Studium: Bedeutsamkeit für den Studienerfolg und Anforderungen an SRL

Auf deskriptiver Ebene zeigte sich zunächst, dass die Expert(inn)en alle einbezogenen typischen Lernsituationen als bedeutsam für den Studienerfolg und als anforderungsreich mit Blick auf die Selbstregulation des Lernens beurteilten (Tab. 2). Darauf verweisen Mittelwerte oberhalb der theoretischen Skalenmitte von 3.5, die für beide

Studienbereich	Prüfungs-vorbereitung	Selbststudium	Vorbereitung eines Vortrags	Erstellung einer Semesterarbeit	Anfertigen einer Ab-schlussarbeit	Teilnahme an einer Vorlesung	Teilnahme an einer Übungs-veranstaltung
			Bedeutsamkeit für den Studienerfolg				
Lehrer(innen)bildung MINT	5.60 (0.79)	5.10 (1.18)	4.07 (1.28)	3.87 (1.26)	4.93 (1.04)	4.57 (1.30)	5.33 (0.96)
Psychologie	5.59 (0.65)	4.72 (1.19)	4.91 (1.12)	4.30 (1.20)	5.63 (0.64)	3.61 (1.33)	4.60 (1.04)
Wirtschaftswissenschaften	5.67 (0.70)	4.99 (1.06)	4.20 (1.29)	4.84 (1.06)	5.47 (0.89)	4.71 (1.23)	4.48 (1.46)
Elektrotechnik	5.48 (0.71)	4.89 (1.03)	4.03 (1.41)	4.41 (1.11)	5.48 (0.62)	4.28 (1.16)	4.97 (1.02)
Gesamt	5.57 (0.71)	4.92 (1.12)	4.33 (1.32)	4.39 (1.20)	5.39 (0.86)	4.29 (1.33)	4.81 (1.21)
			Anforderungen an SRL				
Lehrer(innen)bildung MINT	5.41 (0.73)	5.19 (1.17)	4.63 (1.17)	4.90 (1.02)	5.41 (0.83)	3.87 (1.61)	3.91 (1.44)
Psychologie	5.29 (0.94)	5.13 (1.17)	4.55 (1.20)	5.00 (0.98)	5.79 (0.54)	3.83 (1.47)	4.00 (1.18)
Wirtschaftswissenschaften	5.27 (0.92)	5.24 (0.98)	4.67 (1.13)	4.92 (1.08)	5.56 (0.88)	4.08 (1.45)	3.74 (1.35)
Elektrotechnik	5.25 (0.76)	5.08 (0.86)	4.58 (1.05)	4.81 (1.05)	5.47 (0.73)	3.97 (1.41)	4.37(1.23)
Gesamt	5.30 (0.85)	5.17 (1.05)	4.61 (1.14)	4.92 (1.03)	5.57 (0.77)	3.94 (1.48)	3.98 (1.32)

Anmerkungen. Dargestellt sind Mittelwerte und Standardabweichungen der Expert(inn)enurteile für typische Lernsituationen im Studium von $N = 144$ Dozierenden und $N = 162$ exzellenten Studierenden (Skalenumfang: 1–6).

Tab. 2: Bedeutsamkeit für den Studienerfolg und Anforderungen an SRL von typischen Lernsituationen im Studium, getrennt nach Studienbereichen (dargestellt sind Mittelwerte und Standardabweichungen)

Aspekte in Bezug auf alle sieben Lernsituationen zu beobachten waren. Dies galt für die Gesamtstichprobe, aber auch durchgängig in den Teilstichproben der vier Studienbereiche. Die Korrelationen zwischen den Urteilen zur Bedeutsamkeit der Situationen für den Studienerfolg und deren Anforderungen an das SRL lagen für die verschiedenen Lernsituationen zwischen $r = .29$ und $r = .48$. Die beiden Situationsmerkmale weisen also moderat positive Zusammenhänge auf, sind aber dennoch deutlich voneinander abzugrenzen.

Die Bedeutsamkeit der Lernsituationen für den Studienerfolg sowie ihren Anforderungsgehalt im Hinblick auf die Erfordernis, selbstreguliert zu lernen, wurden inferenzstatistisch jeweils mithilfe von 7 (Lernsituation) × 4 (Studienbereich) faktoriellen Varianzanalysen mit Messwiederholung auf dem ersten Faktor analysiert. In vorgeschalteten Analysen zeigten sich zwischen den beiden Expert(inn)engruppen keine substanziellen Unterschiede, sodass der entsprechende Faktor in die Varianzanalysen nicht einbezogen wurde.[2]

Im Hinblick auf die Bedeutsamkeit der Situationen für den Studienerfolg war ein statistisch signifikanter Haupteffekt der Lernsituation von mittlerer bis großer Effektstärke evident ($F(6,1812) = 79.04$; $p < .001$; $\eta^2 = .21$), der darauf verweist, dass unterschiedliche Lernsituationen in unterschiedlichem Maße wichtig für den Erfolg im Studium sind. Als besonders bedeutsam wurden die Prüfungsvorbereitung und das Erstellen der Abschlussarbeit beurteilt (vgl. Tab. 2). Neben diesem Haupteffekt zeigte sich auch eine Interaktion zwischen Lernsituation und Studienbereich, jedoch mit eher geringer Effektstärke ($F(18,1812) = 10.34$; $p < .001$; $\eta^2 = .09$). Diese Interaktion spiegelt wider, dass unterschiedliche Lernsituationen in unterschiedlichen Studienbereichen in gewissem Maße von unterschiedlicher Bedeutung sind (für Details siehe Tab. 2). Der Haupteffekt des Studienfachs war nicht signifikant ($F(3,302) = 1.27$; $p = .29$; $\eta^2 = .01$).

Bezüglich des SRL-Anforderungsgehalts fanden sich sehr deutliche Unterschiede zwischen den analysierten sieben Lernsituationen (vgl. Tab. 2), indiziert durch einen großen Haupteffekt des Faktors „Lernsituation“ ($F(6,1812) = 117.62$; $p < .001$; $\eta^2 = .28$). Erneut erreichten hier die Anfertigung einer Abschlussarbeit und die Prüfungsvorbereitung sehr hohe Werte, wobei aber die Anforderungen an die Selbstregulation im Zusammenhang des Selbststudiums, der Erstellung einer Semesterarbeit oder auch der Vorbereitung eines Vortrags von den Expert(inn)en durchaus ähnlich groß beurteilt

2 Diese Vorabanalysen bestanden aus den dargestellten Varianzanalysen, die um den zweistufigen Faktor der Expert(inn)engruppe erweitert wurden. Weder bei der Bedeutsamkeit für den Studienerfolg noch bei den SRL-Anforderungen fand sich ein signifikanter Haupteffekt der Expert(inn)engruppe ($p > .08$). Auch die jeweils drei getesteten Interaktionen mit der Expert(inn)engruppe erwiesen sich als nicht signifikant ($p > .13$). Die einzige Ausnahme davon stellte eine statistisch signifikante Interaktion zwischen Lernsituation und Expert(inn)engruppe bei den SRL-Anforderungen dar, die allerdings von geringer Effektstärke und damit praktisch wenig bedeutsam war ($p = .03$; $\eta^2 = .01$) – kleinere Unterschiede bestanden hier bei den Beurteilungen der Erstellung von Semesterarbeiten, der Teilnahme an Übungsveranstaltungen oder der Teilnahme an Vorlesungen, wobei die studentischen Expert(inn)en die Anforderungen an SRL jeweils höher einschätzten als die Dozierenden.

wurden. Interessanterweise zeigten sich in Bezug auf die SRL-Anforderungen keinerlei Fachspezifika: Weder der Haupteffekt des Studienbereichs ($F(3,302) = 0.05$; $p = .98$; $\eta^2 = .00$) noch dessen Interaktion mit der Lernsituation waren signifikant ($F(18,1812) = 1.58$; $p = .06$; $\eta^2 = .02$).

3.2 Strategiepassung in verschiedenen Lernsituationen

Die Passung der theoretisch abgeleiteten SRL-Strategien zu den einbezogenen typischen Lernsituationen wurde mithilfe einer weiteren Varianzanalyse analysiert, nämlich einer 7 (Lernsituation) × 4 (Studienbereich) × 9 (SRL-Strategie) faktoriellen Varianzanalyse mit Messwiederholung auf dem ersten und dritten Faktor. Erneut erbrachte eine vorgeschaltete Analyse einen großen Konsens der beiden Gruppen an Praxisexpert(inn)en, sodass dieser Faktor nicht weiter berücksichtigt wurde.[3]

Von sehr großer Effektstärke waren die signifikanten Haupteffekte der Lernsituation ($F(6,1812) = 112.54$; $p < .001$; $\eta^2 = .27$) und der SRL-Strategie ($F(6,2416) = 148.80$; $p < .001$; $\eta^2 = .33$). Ersterer spiegelte die Ergebnisse zu den Anforderungen an SRL sehr gut wider (allgemein größte Bedeutung von SRL-Strategien bei der Prüfungsvorbereitung und der Anfertigung von Abschlussarbeiten, geringste Bedeutung bei der Teilnahme an Vorlesungen und Übungsveranstaltungen). Der Haupteffekt der SRL-Strategie bestätigte eine Grundannahme des vorliegenden Ansatzes, dass nämlich nicht alle SRL-Strategien von gleicher Bedeutung sind und entsprechend eine Maximalausprägung aller Strategien wenig funktional ist.

Wesentlich im Hinblick auf den grundlegenden Passungsgedanken ist, dass die Expert(inn)en unterschiedliche SRL-Strategien in unterschiedlichen Lernsituationen als unterschiedlich geeignet einschätzten – der entsprechende Interaktionseffekt zwischen den beiden Faktoren „Lernsituation“ und „SRL-Strategie“ wies eine mittlere bis große Effektstärke auf ($F(48,14496) = 63.80$; $p < .001$; $\eta^2 = .17$). Dies indiziert, dass die Eignungen verschiedener SRL-Strategien in Abhängigkeit von der Lernsituation variieren. Die situationsspezifischen Eignungen der einbezogenen SRL-Strategien sind in Abbildung 2 dargestellt. Zwei Beispiele für diese situationsspezifischen Eignungen sind: (1) Oberflächenstrategien (z. B. Auswendiglernen) werden bei der Prüfungsvorbereitung als vergleichsweise wichtig eingeschätzt, während sie in Situationen wie der Anfertigung von Abschlussarbeiten oder der Vorbereitung von Vorträgen als recht unangemessen beurteilt werden. (2) Das Management externer Ressourcen (wie das Einholen von Hilfe) wird für Situationen, in denen überwiegend eigenständiges Arbeiten erforderlich ist (z. B. Erstellung wissenschaftlicher Arbeiten, Prüfungsvorbereitung, Selbst-

3 Diese äquivalent zu oben durchgeführten Vorabanalysen erbrachten einen nicht-signifikanten Haupteffekt der Expert(inn)engruppe ($p = .45$) sowie aufgrund geringer Effektstärken durchwegs zwar statistisch, jedoch nicht praktisch bedeutsame Interaktionseffekte mit diesem Faktor ($p < .05$; $\eta^2 < .03$).

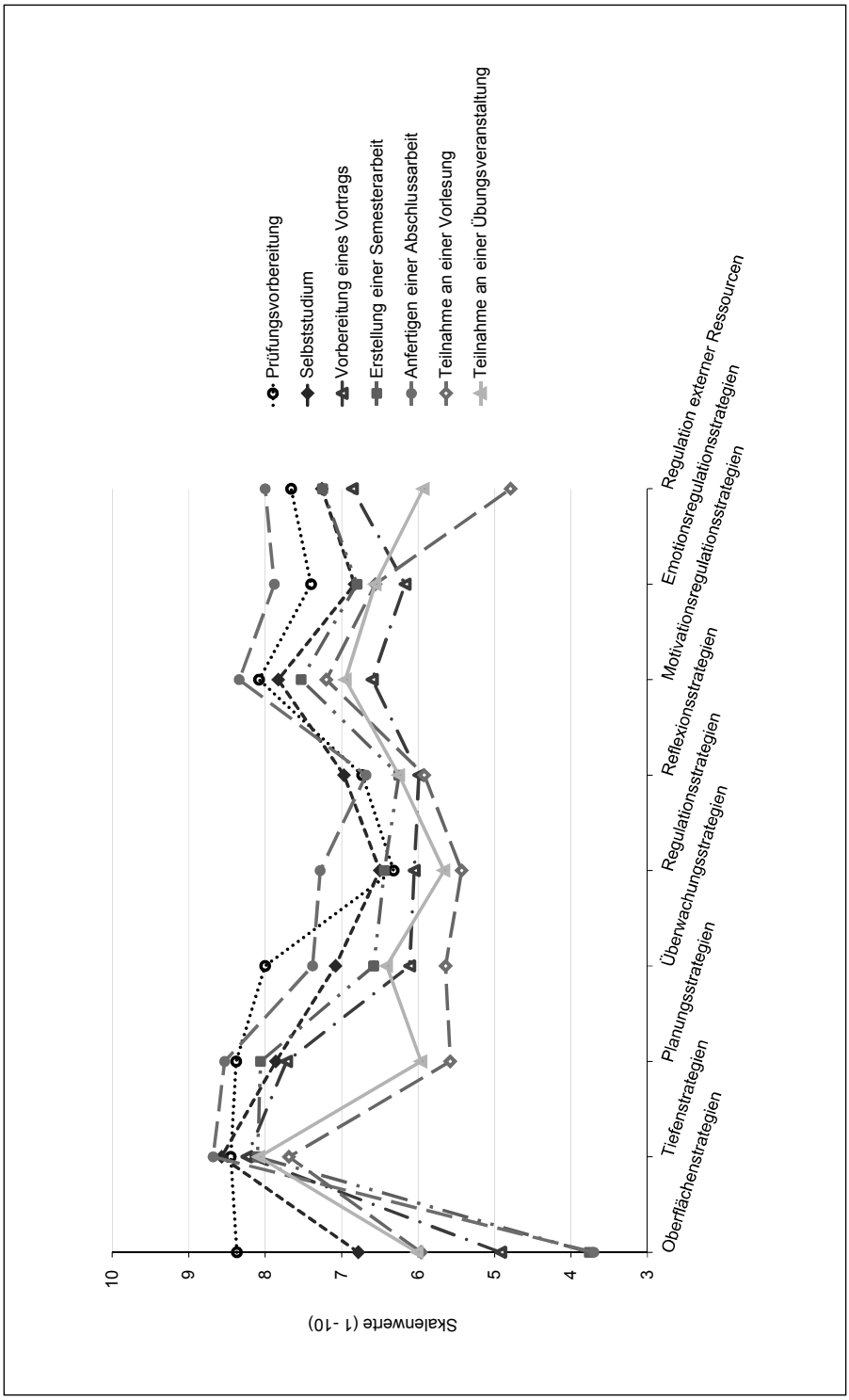


Abb. 2: Eignung von neun theoretisch abgeleiteten Gruppen von SRL-Strategien für sieben typische Lernsituationen im Studium (dargestellt sind Mittelwerte aller Expert(inn)en)

studium), als wesentlich wichtiger eingeschätzt als bei zeitlich und räumlich vorstrukturierten Lernsituationen (z. B. Besuch von Vorlesungen oder Übungsveranstaltungen).

In Bezug auf mögliche Fachspezifika der Strategieeignung zeigten sich ein Haupteffekt des Studienbereichs ($F(3,298) = 5.03$; $p < .01$; $\eta^2 = .05$), Interaktionen zwischen Studienbereich einerseits und Lernsituation ($F(18,1812) = 2.80$; $p < .001$; $\eta^2 = .03$) und SRL-Strategie ($F(24,2416) = 2.15$; $p < .01$; $\eta^2 = .02$) andererseits sowie eine Dreifach-Interaktion zwischen allen drei Faktoren ($F(144,14496) = 2.13$; $p < .001$; $\eta^2 = .02$). Alle Effekte indizieren, dass die Strategieeignungsurteile der Expert(inn)en von ihrer Fachzugehörigkeit abhingen (z. B. variierte die eingeschätzte Passung zwischen Strategie und Situation zwischen Studienbereichen). Sie waren allerdings allesamt von eher geringer Stärke, was auf insgesamt allenfalls moderate Fachspezifika verweist.

Um aus den zusätzlichen Prioritätswahlen der Expert(inn)en (Identifikation der drei wichtigsten SRL-Strategien für jede Situation) erste Hinweise zur Validität der im Arbeitsmodell angenommenen Kombination von Strategie- und Prozessdimension abzuleiten, wurden die Häufigkeiten ermittelt, mit denen die Expert(inn)en in den einzelnen Situationen die einbezogenen Strategien als besonders wichtig eingestuft hatten. Basierend auf den in Abschnitt 2.2 dargestellten Zuordnungen wurden dazu die Nennungshäufigkeiten für alle theoretisch plausiblen Zellen bestimmt, die sich aus der Kombination der beiden Dimensionen ergeben (Abb. 3). Strategien, die nicht eindeutig bestimmten Phasen des Lernprozesses zuzuordnen sind (z. B. Strategien zur Regulation der eigenen Motivation, die in verschiedenen Phasen angewendet werden können; vgl. Engelschalk, Steuer & Dresel, im Druck), wurden dabei für alle betreffenden Zellen gewertet.

Substanzielle Nennungshäufigkeiten ergaben sich für alle sieben Zellen, was dafür spricht, dass sich aus der Kombination von Strategie- und Prozessdimension bedeutsame Facetten von SRL-Strategien ableiten lassen. Die Zellen unterschieden sich jedoch deutlich in den Nennungshäufigkeiten. Beispielsweise wurden SRL-Strategien, die in der postaktionalen Phase verortet sind, insgesamt als deutlich weniger bedeutsam beurteilt als Strategien der präaktionalen und aktionalen Phase. Zudem ergaben sich Unterschiede zwischen den einzelnen Situationen, die wiederum im Einklang mit dem grundlegenden Passungsgedanken standen. So lagen die Nennungshäufigkeiten für kognitive Strategien bei stärker strukturierten Lernsituationen (z. B. Teilnahme an einer Übungsveranstaltung oder einer Vorlesung) und klassischen Selbstlernsituationen (Selbststudium, Prüfungsvorbereitung) recht hoch, während sie für das Anfertigen von Abschluss- oder Semesterarbeiten (Erledigung selbstorganisierter, komplexer Anforderungen) weniger häufig als wichtig beurteilt wurden. In diesen komplexen Anforderungssituationen wurden hingegen metakognitive Strategien häufiger als zentral erachtet. Ressourcenbezogene Strategien wurden übergreifend über alle Situationen vergleichsweise häufig als wichtige Strategien gewählt.

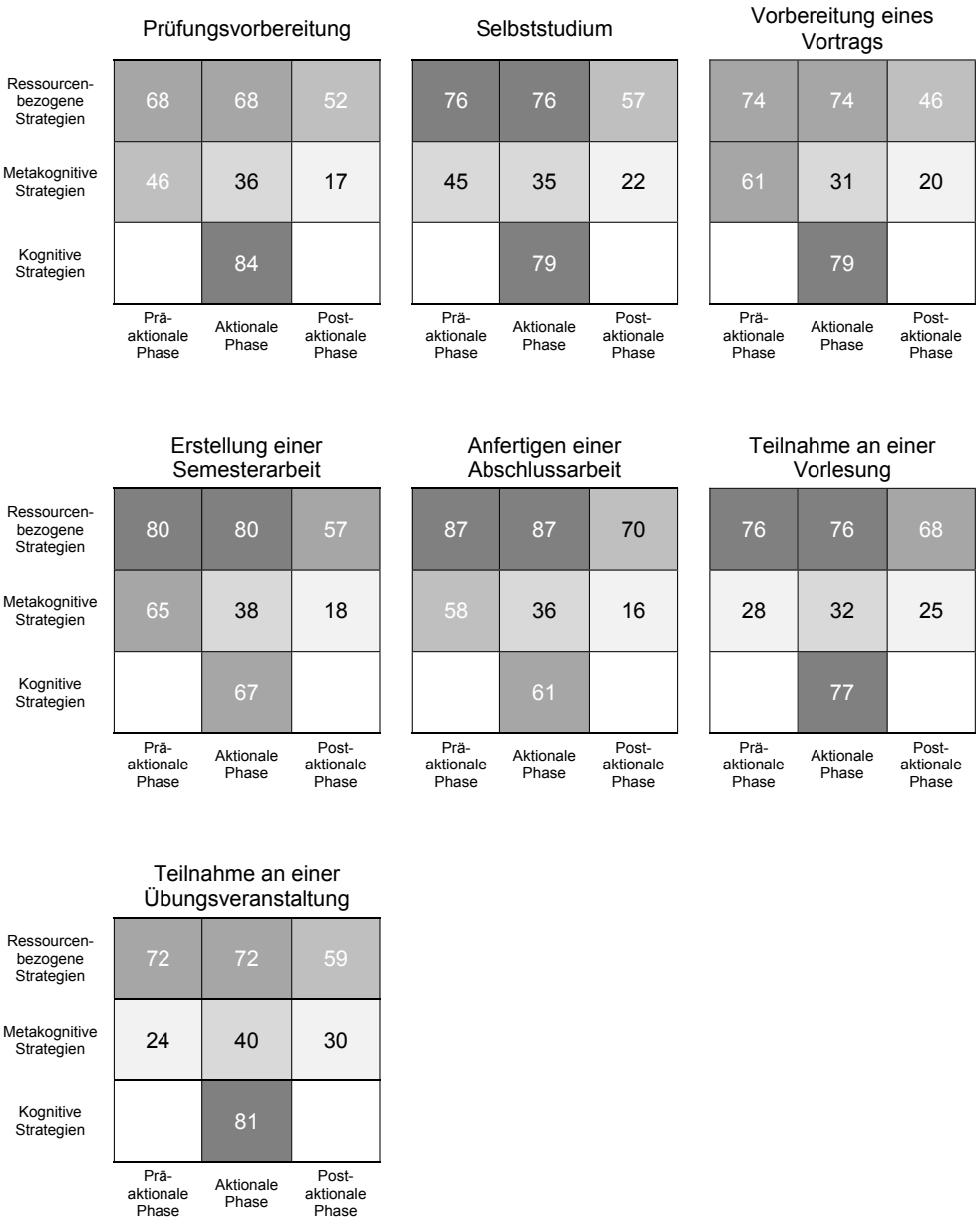


Abb. 3: Häufigkeiten, mit denen SRL-Strategien in sieben theoretisch begründbaren Kombinationen aus Strategie- und Prozessdimension des vorgeschlagenen Arbeitsmodells von Expert(inn)en als besonders geeignet klassifiziert wurden (alle Angaben in %)

4. Diskussion

Ziel dieses Beitrags war es, belastbare quantitative Evidenzen zum einen zur relativen Bedeutung sowie den Anforderungen unterschiedlicher Lernsituationen in unterschiedlichen Studienbereichen und, zum anderen, zur relativen Eignung unterschiedlicher SRL-Strategien in diesen Situationen zu generieren. Die untersuchten Strategien wurden dabei aus einem differenziert und umfassend konzipierten Arbeitsmodell abgeleitet, das einen integrativen Ansatz zur Konzeption und Struktur von SRL-Kompetenzen im tertiären Bereich liefert.

Die vorliegende quantitative Arbeit setzte dabei an den Ergebnissen einer qualitativen Vorarbeit an (Dresel et al., im Druck), mit der sie gemeinsam ein sequenzielles Mixed-Method-Vorgehen und damit ein starkes methodisches Design bildet. Die weitgehende Bestätigung der Ergebnisse der vorangegangenen qualitativen Untersuchung durch die vorliegenden Ergebnisse kann als Hinweis auf die Gültigkeit der Ergebnisse insgesamt gewertet werden. Eine weitere Stärke der vorliegenden Arbeit besteht in dem Fokus auf Expert(inn)en für die Praxis des SRL im tertiären Bereich, durch den die theoretischen Perspektiven der SRL-Literatur (für einen Überblick siehe z. B. Zimmerman & Schunk, 2011) sinnvoll ergänzt werden können. Jenßen et al. (2015, in diesem Beiheft) weisen für den Kontext der Messinstrumenteentwicklung, in dem die vorliegende Arbeit eingebettet ist, auf die Notwendigkeit der frühen Einbindung von Expert(inn)en hin, um die Inhaltsvalidität zu sichern. Dies wurde durch den Einbezug zweier verschiedener Gruppen an Expert(inn)en für die Praxis des SRL im Hochschulbereich (für das Lehren und Lernen in ihrem Studienbereich) realisiert, wodurch das Risiko von Urteilsfehlern reduziert wurde. Die große Übereinstimmung, die beide Expert(inn)engruppen in ihren Einschätzungen hatten, spricht dabei für die Belastbarkeit der Ergebnisse. Schließlich liegen Stärken der vorliegenden Arbeit im intensiven Fokus auf spezifische Lernsituationen, der aktuellen theoretischen Ansätzen in der SRL-Forschung entspricht (z. B. Winne, 2010; Wirth & Leutner, 2008), sowie im Einbezug von vier inhaltlich und strukturell recht unterschiedlichen Studienbereichen, die jeweils an zwei verschiedenen Hochschulstandorten untersucht wurden, was der Generalisierbarkeit der Befunde zugute kommt.

Die Ergebnisse legen nahe, dass über alle Fächer hinweg eine stabile Kerngruppe an Lernsituationen existiert, die eine hohe Relevanz für den Studienerfolg haben und die fachübergreifend anforderungsreich im Hinblick auf die Selbstregulation des Lernens Studierender sind. Dabei waren die unterschiedlichen Lernsituationen im Expert(inn)enurteil wie erwartet von heterogener Relevanz als auch von heterogenem Anforderungsgehalt. Typische Selbstlernsituationen sowie Aufgaben mit komplexen Anforderungen und hohem Selbstorganisationsaufwand wurden dabei als besonders bedeutsam für den Studienerfolg und auch als besonders anforderungsintensiv in Bezug auf SRL beurteilt. Situationen, die den Besuch sowie die Vor- und Nachbereitung mehr oder weniger strukturierter Lerngelegenheiten betreffen, wurden insbesondere hinsichtlich ihrer Anforderungen an SRL geringer eingeschätzt (z. B. Teilnahme an einer Vorlesung oder einer Übungsveranstaltung). Diese von Lernsituation zu Lernsituation variierenden Anforderungen liefern empirische Evidenz für die in neueren theoretischen Arbei-

ten geforderte Berücksichtigung von multiplen und spezifischen Situationen bei der Untersuchung von Selbstregulationsprozessen beim Lernen (vgl. Winne, 2010). Festzuhalten bleibt bei allen erwarteten Unterschieden jedoch auch, dass alle sieben analysierten Lernsituationen insgesamt als hinreichend bedeutsam für den Studienerfolg und als hinreichend mit Anforderungen an die Selbstregulation ausgestattet bewertet wurden. Dieses Befundmuster galt für alle vier einbezogenen Studienbereiche und war bei beiden Expert(inn)engruppen evident – dies legt eine Generalisierbarkeit dieser Ergebnisse auch darüber hinaus nahe.

Die vorliegende Studie erbrachte auf der Basis der Expertise von Praktiker(inne)n des SRL Hinweise darauf, dass Strategien aller drei makroskopischer Arten (kognitive Lernstrategien, metakognitive Strategien, Strategien des Ressourcenmanagements) bedeutsam für das Lernen im tertiären Bereich sind (Strategiedimension). Zudem ergaben sich auch Hinweise, dass SRL-Strategien sowohl in allen Phasen des Lernprozesses (Prozessdimension) als auch in allen theoretisch plausiblen Kombinationen beider Dimensionen bedeutsam sind.

Damit stehen die gewonnenen Befunde im Einklang mit dem vorgeschlagenen Arbeitsmodell und liefern ein Indiz für dessen Gültigkeit. Gleichwohl muss angemerkt werden, dass eine belastbare Überprüfung der Modellstruktur nur anhand der Ergebnisse von Studierenden in entsprechenden (erst noch zu entwickelnden) Kompetenztests etwa mit faktorenanalytischen Methoden erfolgen kann. Damit müssen zukünftige Untersuchungen erst erweisen, ob die theoretisch ausdifferenzierte Modellstruktur empirisch haltbar ist. Dennoch können die zentralen Stärken des vorgeschlagenen Modells, nämlich der große Differenziertheitsgrad und der umfassende theoretische Fokus (die beide aus der eklektischen Anlage, d. h. der Kombination verschiedener theoretischer Perspektiven resultieren), bereits ihre Wirkung entfalten, wenn ein breites theoretisches Suchraster erforderlich ist – etwa wenn blinde Flecken der SRL-Forschung identifiziert, umfassende Intervention zur SRL-Förderung konzipiert oder differenzierte Messinstrumente entwickelt werden sollen. Letztlich ist es aber nötig, sich zu vergegenwärtigen, dass das vorgeschlagene Arbeitsmodell nicht für alle Forschungszwecke geeignet sein kann und die Entscheidung für einen bestimmten Grad an Modellkomplexität stets bewusst und in engem Zusammenhang mit dem Forschungsanliegen erfolgen sollte.

Zentral in Bezug auf die Passungsannahme und die damit im Zusammenhang stehende Anforderung, qualitative Standards zur Beurteilung von SRL-Kompetenzen zu definieren, ist das Urteil der Expert(inn)en, dass in verschiedenen relevanten Lernsituationen unterschiedliche Gruppen von SRL-Strategien angemessen sind – dies zeigte sich übereinstimmend sowohl anhand der mit Ratingskalen erfassten Urteile als auch anhand der Prioritätswahlen. So ergaben sich Hinweise etwa darauf, dass kognitive Strategien in Situationen wie der Vor- und Nachbereitung von Lehrveranstaltungen sowie in der Prüfungsvorbereitung angemessen sind, diese jedoch kaum geeignet sind, wenn es um komplexere Lernsituationen geht, die höhere Anforderungen an die Selbstregulation stellen (wie beispielsweise das Erstellen einer wissenschaftlichen Arbeit). Metakognitive Strategien (v. a. Planung und Überwachung) und Strategien des Ressourcenmanagements wurden als besonders geeignet für das Erstellen wissenschaftlicher

Arbeiten und die Prüfungsvorbereitung angesehen, kaum jedoch für stark von außen strukturierte Situationen (z. B. Vorlesungsbesuch). Beachtenswert ist, dass insbesondere metakognitive Strategien (in der Tendenz aber auch ressourcenbezogene Strategien) in der postaktionalen Phase als weniger wichtig eingeschätzt wurden. In Prozessmodellen des SRL wird diese Phase dagegen als gleichwertig zu den beiden anderen Phasen konzeptualisiert (vgl. Schmitz & Wiese, 2006; Zimmerman, 2000). Insgesamt stehen die Ergebnisse im Einklang mit der theoretischen Annahme, dass eine maximale Nutzung aller SRL-Strategien in allen Situationen weniger funktional ist als ein an die spezifischen Anforderungen angepasster Strategieeinsatz (vgl. Wirth & Leutner, 2008).

Auf Basis der Ergebnisse wird deutlich, dass eine situationsunspezifische Konzeption von SRL-Kompetenzen kaum valide sein kann, da zum einen die Anforderungen an SRL und zum anderen auch die idealerweise einzusetzenden Strategien situativ variieren. Daher ist es angezeigt, Charakteristika der Lernsituation deutlich stärker als bisher in der Theoriebildung üblich zu berücksichtigen und eine grundsätzlich situationsspezifische Herangehensweise zu realisieren (vgl. Winne, 2010).

Die vorliegende Studie weist trotz ihrer oben ausgeführten Stärken einige Limitationen auf, die bei der Ergebnisinterpretation beachtet werden sollten. Dies betrifft zunächst die Stichproben, die bis zu einem gewissen Grad für die einzelnen Fächer unbalanciert waren (da an einigen Standorten Stichprobenausfälle zu verzeichnen waren). Allerdings sind die Unterschiede zwischen den Substichproben eher klein und wurden in den gewählten statistischen Ansätzen kontrolliert. Des Weiteren wurde in der vorliegenden Arbeit die Differenzierung nach verschiedenen Arten von Strategiewissen (Wissensdimension im vorgeschlagenen Arbeitsmodell) nicht explizit berücksichtigt. Da nach der Passung von Strategien in bestimmten Situationen gefragt wurde, wurden Evidenzen gewonnen, die primär zur Beurteilung von konditionalem Strategiewissen herangezogen werden können – dieses Vorgehen wurde gewählt, da diese Wissensart als besonders relevant für die Erfassung von Kompetenzen zum SRL gelten kann (vgl. Dresel et al., im Druck; Paris et al., 1983). Dieses Wissen setzt zwar entsprechendes deklaratives und prozedurales Wissen voraus, dennoch sind auf Basis der vorliegenden Studie noch keine verlässlichen Aussagen über deren Bedeutung möglich. Die Untersuchung der unterschiedlichen Wissensarten und deren Erklärungswert für Outcomevariablen sollte Gegenstand zukünftiger Studien sein.

Im Sinne einer Weiterführung der vorliegenden Arbeit sind darüber hinaus insbesondere Neuentwicklungen von Messinstrumenten zur Erfassung von SRL angezeigt, um die bezüglich ihrer Validität noch bestehenden Limitationen zu überwinden. Naheliegend erscheinen auf der Grundlage der hier vorgestellten Ergebnisse vor allem situationsspezifische Messinstrumente, die ggf. auch zu mehr als nur einem Messzeitpunkt vorgelegt werden. Die Situationsvignetten bieten konkrete Beschreibungen bedeutsamer Anforderungssituationen im Studium, die beispielsweise in Situational Judgment Tests Verwendung finden und in einer validen Erfassung von SRL-Kompetenzen münden könnten. Die Studie bildet damit eine solide Basis für zukünftige Studien sowie die Konstruktion situationsbasierter Messinstrumente zur Erfassung von SRL im tertiären Kontext.

Literatur

Artelt, C. (2000). *Strategisches Lernen*. Münster: Waxmann.

Artelt, C., Demmrich, A., & Baumert, J. (2001). Selbstreguliertes Lernen. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 271–298). Opladen: Leske + Budrich.

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research, 31*, 445–457.

Boekaerts, M., & Corno, L. (2005). Self-Regulation in the Classroom: A Perspective on Assessment and Intervention. *Applied Psychology: An International Review, 54*, 199–231.

Boekaerts, M., Pintrich, P., & Zeidner, M. (Hrsg.) (2000). *Handbook of self-regulation*. Orlando: Academic Press.

Dresel, M., & Haugwitz, M. (2005). The relationship between cognitive abilities and self-regulated learning: Evidence for interactions with academic self-concept and gender. *High Ability Studies, 16*, 201–218.

Dresel, M., Schmitz, B., Schober, B., Spiel, C., Ziegler, A., Engelschalk, T., Jöstl, G., Klug, J., Roth, A., Wimmer, B., & Steuer, G. (im Druck). Competencies for successful self-regulated learning in higher education: Structural model and empirical evidence from expert interviews. *Studies in Higher Education*.

Engelschalk, T., Steuer, G., & Dresel, M. (im Druck). Wie spezifisch regulieren Studierende ihre Motivation bei unterschiedlichen Anlässen? Ergebnisse einer Interviewstudie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*.

Friedrich, H. F., & Mandl, H. (1995). Analyse und Förderung selbstgesteuerten Lernens. In F. Weinert & H. Mandl (Hrsg.), *Psychologie der Erwachsenenbildung* (Enzyklopädie der Psychologie, Serie Pädagogische Psychologie, Bd. 4, S. 237–293). Göttingen: Hogrefe.

Händel, M., Artelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal for Educational Research Online, 5*, 162–188.

Heckhausen, H., & Gollwitzer, P. M. (1987). Thought contents and cognitive functioning in motivational versus volitional states of mind. *Motivation and Emotion, 11*, 101–120.

Jenßen, L., Dunekacke, S., & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik, 61. Beiheft*, 11–31.

Lehmann, M., & Hasselhorn, M. (2009). Entwicklung von Lernstrategien im Grundschulalter. In F. Hellmich & S. Wernke (Hrsg.), *Lernstrategien im Grundschulalter* (S. 25–41). Stuttgart: Kohlhammer.

Leutner, D., Barthel, A., & Schreiber, B. (2001). Studierende können lernen, sich selbst zum Lernen zu motivieren: Ein Trainingsexperiment. *Zeitschrift für Pädagogische Psychologie, 15*, 155–167.

Marton, F., & Saljö, R. (1984). Approaches to learning. In F. Marton, D. J. Hounsell & N. J. Entwistle (Hrsg.), *The experience of learning* (S. 36–55). Edinburgh: Scottish Academic Press.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic learner. *Contemporary Educational Psychology, 8*, 293–316.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist, 37*, 91–106.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* Ann Arbor: University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning.

Schiefele, U., & Pekrun, R. (1996). Psychologische Modelle des fremdgesteuerten und selbstgesteuerten Lernens. In F. E. Weinert (Hrsg.), *Psychologie des Lernens und der Instruktion*

(Enzyklopädie der Psychologie, Serie Pädagogische Psychologie, Bd. 2, S. 249–278). Göttingen: Hogrefe.

Schlagmüller, M., & Schneider, W. (2007). *WLST 7–12. Würzburger Lesestrategie-Wissenstest für die Klassen 7 bis 12*. Göttingen: Hogrefe.

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology, 31*, 64–96.

Schneider, W., & Lockl, K. (2006). Entwicklung metakognitiver Kompetenzen im Kindes- und Jugendalter. In W. Schneider & B. Sodian (Hrsg.), *Kognitive Entwicklung* (S. 721–767). Göttingen: Hogrefe.

Schober, B., Klug, J., Spiel, C., Dresel, M., Steuer, G., Schmitz, B., & Ziegler, A. (im Druck). Gaining substantial new insights into university students' SRL competences – What do we need to succeed? *Journal of Psychology/Zeitschrift für Psychologie*.

Schunk, D. H., & Zimmerman, B. J. (2003). Self-regulation and learning. In W. M. Reynolds & G. E. Miller (Hrsg.), *Handbook of psychology: Educational psychology* (S. 59–78). New York: Wiley.

Schwinger, M., Steinmayr, R., & Spinath, B. (2009). How do motivational regulation strategies affect achievement: Mediated by effort management and moderated by intelligence. *Learning and Individual Differences, 19*, 621–627.

Shraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*, 351–371.

Spörer, N., & Brunstein, J. C. (2006). Erfassung selbstregulierten Lernens mit Selbstberichtsverfahren: Ein Überblick zum Stand der Forschung. *Zeitschrift für Pädagogische Psychologie, 20, 147–160.*

Ständige Kultusministerkonferenz der Länder (2005). *Qualifikationsrahmen für Deutsche Hochschulabschlüsse*.http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2005/2005_04_21-Qualifikationsrahmen-HS-Abschluesse.pdf [15. 05. 2014].

Thillmann, H. (2007). *Selbstreguliertes Lernen durch Experimentieren: Von der Erfassung zur Förderung* (Dissertation). Universität Duisburg-Essen.

Weinstein, C. E., & Hume, L. M. (1998). *Study strategies for lifelong learning*. Washington, D. C.: American Psychological Association.

Wild, K.-P., & Schiefele, U. (1994). Lernstrategien im Studium. Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens. *Zeitschrift für Differentielle und Diagnostische Psychologie, 15*, 185–200.

Wild, K.-P., Schiefele, U., & Winteler, A. (1992). *LIST – Ein Verfahren zur Erfassung von Lernstrategien im Studium*. Neubiberg: Gelbe Reihe.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist, 45*, 267–276.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Hrsg.), *Metacognition in educational theory and practice* (S. 277–304). Mahwah: Lawrence Erlbaum Associates.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. Pintrich & M. Zeidner (Hrsg.), *Handbook of self-regulation* (S. 531–566). Orlando: Academic Press.

Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence: Implications of theoretical models for assessment methods. *Zeitschrift für Psychologie/Journal of Psychology, 216*, 102–110.

Wolters, C. A. (2003). Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning. *Educational Psychologist, 38*, 189–205.

Zauchner, S., Baumgartner, P., Blaschitz, E., & Weissenback, A. (Hrsg.) (2008). *Offener Bildungsraum Hochschule: Freiheiten und Notwendigkeiten*. Münster: Waxmann.
Ziegler, A., Porath, M., & Stöger, H. (Gast-Hrsg.) (2011). Quantitative approaches to the study of self-regulated learning. *Psychological Test and Assessment Modeling* (Special Issue).
Ziegler, A., Stöger, H., & Dresel, M. (2004). Selbstreguliertes Lernen. In C. A. von Gleichenstein (Hrsg.), *Schulpsychologie als Brücke zwischen Familie und Schule* (S. 23–32). Bonn: Deutscher Psychologen-Verlag.
Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Hrsg.), *Handbook of self-regulation* (S. 13–39). San Diego: Academic Press.
Zimmerman, B. J., & Schunk, D. (Hrsg.) (2011). *Handbook of self-regulation of learning and performance*. New York: Routledge.

Abstract: Competencies in self-regulated learning (SRL) are seen as central prerequisites for success in academic studies. It is still unclear as to which SRL strategies are most effective in which learning situations. Based on surveys conducted among 306 expert SRL practitioners, the present work provides evidence regarding the relative importance, and SRL demands, of different learning situations in four fields of study. Furthermore, expert assessments regarding the relative fit of different SRL strategies in these situations are reported. The strategies examined here were derived from a differentiated and comprehensively conceptualized working model that provides an integrative approach to the design and structure of SRL competencies in higher education.

Keywords: Self-Regulated Learning, Assessment of Competencies, Learning Situations, Learning Strategies, Expert Survey

Anschrift der Autor(inn)en

Dr. Gabriele Steuer, Universität Augsburg, Lehrstuhl für Psychologie,
Universitätsstraße 10, 86159 Augsburg, Deutschland
E-Mail: gabriele.steuer@phil.uni-augsburg.de

Tobias Engelschalk, Universität Augsburg, Lehrstuhl für Psychologie,
Universitätsstraße 10, 86159 Augsburg, Deutschland
E-Mail: tobias.engelschalk@phil.uni-augsburg.de

Mag. Gregor Jöstl, Universität Wien, Arbeitsbereich Bildungspsychologie
und Evaluation, Universitätsstraße 7 (NIG), 1010 Wien, Österreich
E-Mail: gregor.joestl@univie.ac.at

Dipl. Psych. Anne Roth, Technische Universität Darmstadt, Lehrstuhl für
Pädagogische Psychologie, Alexanderstraße 10, 64283 Darmstadt, Deutschland
E-Mail: roth@psychologie.tu-darmstadt.de

Bastian Wimmer, Universität Erlangen-Nürnberg, Lehrstuhl für Pädagogische Psychologie,
Regensburger Straße 160, 90478 Nürnberg, Deutschland
E-Mail: bastian.wimmer@fau.de

Prof. Dr. Bernhard Schmitz, Technische Universität Darmstadt,
Lehrstuhl für Pädagogische Psychologie, Alexanderstraße 10, 64283 Darmstadt, Deutschland
E-Mail: schmitz@psychologie.tu-darmstadt.de

Prof. Dr. Barbara Schober, Universität Wien, Arbeitsbereich Bildungspsychologie
und Evaluation, Universitätsstraße 7 (NIG), 1010 Wien, Österreich
E-Mail: barbara.schober@univie.ac.at

Prof. Drs. Christiane Spiel, Universität Wien, Arbeitsbereich Bildungspsychologie
und Evaluation, Universitätsstraße 7 (NIG), 1010 Wien, Österreich
E-Mail: christiane.spiel@univie.ac.at

Prof. Dr. Albert Ziegler, Universität Erlangen-Nürnberg, Lehrstuhl für Pädagogische
Psychologie, Regensburger Straße 160, 90478 Nürnberg, Deutschland
E-Mail: albert.ziegler@fau.de

Prof. Dr. Markus Dresel, Universität Augsburg, Lehrstuhl für Psychologie,
Universitätsstraße 10, 86159 Augsburg, Deutschland
E-Mail: markus.dresel@phil.uni-augsburg.de

Johannes König

Stand der Forschung zu wissenschaftsbezogenen Kompetenzen und weiterführende Fragen

Ein Kommentar

1. Einleitung

Der Kompetenzerwerb von Studierenden im Hochschulbereich steht in einem Spannungsverhältnis zwischen der Orientierung an der jeweiligen wissenschaftlichen Fachdisziplin und einer Orientierung an beruflichen Anforderungen. Erwartungen an den berufsausbildungsnahen Charakter von Studiengängen, wie sie mit der Einführung der neuen Studienstrukturen vielfach formuliert werden (u. a. bereits Wissenschaftsrat, 2000), ziehen die Diskussion um den Erwerb berufsbezogener Kompetenzen nach sich. Die Betonung der Wichtigkeit einer wissenschaftlichen Ausbildung auch für Studiengänge mit konkretem Berufsziel (z. B. Lehramt) führt dagegen zu der Notwendigkeit, wissenschaftsbezogene Kompetenzen in den Blick zu nehmen. Im Zuge des Wandels von Studienbedingungen wie zeitliche Verkürzung, Erhöhung der Studierendenzahlen oder Modularisierung und Straffung des Curriculums, aber auch im Zuge gestiegener Erwartungen an die Stärkung professioneller Praxen durch wissenschaftliche Erkenntnisse – vermittelt über die Studienqualifikationen der beteiligten Akteure – rücken Fragen der Qualitätssicherung im Hochschulbereich stärker als bislang in den Vordergrund.

Die hochschulbezogene Kompetenzforschung steht somit vor der Herausforderung, wissenschaftsbezogene Kompetenzen zu konzeptualisieren und einer empirischen Messung zugänglich zu machen, damit weiterführende Fragestellungen in empirischen Studien zum Erwerb wissenschaftsbezogener Kompetenzen im Hochschulsektor sowie zur Wirksamkeit von Studiengängen an Hochschulen überhaupt erst bearbeitet werden können. Bilanzierende Ergebnisse solcher Studien dienen dann z. B. dazu, Lernbedarfe Studierender zu diagnostizieren, curriculare Entscheidungen zu begründen oder *good practice* zu identifizieren. Experimentelle Studien könnten mithilfe der entwickelten Instrumente die Lernwirksamkeit neuer curricularer Elemente analysieren.

Die in diesem Teil des vorliegenden Beiheftes der *Zeitschrift für Pädagogik* versammelten und im Folgenden zu kommentierenden Beiträge nehmen sich der an die hochschulbezogene Kompetenzforschung gestellten Herausforderung an, indem sie wissenschaftsbezogene Kompetenzen konzeptualisieren und messen. Einerseits kann die Aktualität ihres Anliegens damit kaum hoch genug eingeschätzt werden, andererseits ist mit ihnen die Erwartung verbunden, Grundlagen für weiterführende Forschung zu entwickeln und zu etablieren – also innovativ zu sein und nach Möglichkeit Standards zu

setzen. Zugleich stehen sie vor dem Hintergrund der bisherigen Forschung zur Kompetenzmessung im tertiären Bildungsbereich, wie sie etwa von der *International Association for the Evaluation of Educational Achievement* (IEA) initiiert und als „neuer Typus des international vergleichenden large scale assessments" (Klieme, 2012, S. 492) identifiziert wurde. Somit soll im Folgenden auf die vier Beiträge bilanzierend eingegangen werden und ihre Bezüge zur bereits vorliegenden Kompetenzmessung im tertiären Bildungsbereich – am Beispiel der IEA-Studie *Teacher Education and Development Study: Learning to Teach Mathematics* (TEDS-M; Blömeke, Kaiser & Lehmann, 2010a, 2010b; Tatto et al., 2012) – aufgezeigt werden, um unter anderem vor diesem Hintergrund weiterführende Fragen zu entwickeln.

TEDS-M ist die bislang größte internationale Vergleichsstudie zur Messung der professionellen Kompetenzen angehender Mathematiklehrkräfte im letzten Jahr ihrer Ausbildung. Kern der Studie war die Testung von drei Wissenskonstrukten – mathematisches Fachwissen, pädagogisches Wissen, mathematikdidaktisches Wissen –, die die drei zentralen Kategorien des Lehrerwissens darstellen (Shulman, 1987) und als kognitive Elemente professioneller Kompetenz von Lehrkräften betrachtet werden (Baumert & Kunter, 2006). Mithilfe repräsentativer Länderstichproben wurden bilanzierende Aussagen zu Ergebnissen der Lehrerbildungssysteme getroffen.

2. Fragen an die Beiträge

Die vier Beiträge, die den Teil der wissenschaftsbezogenen Kompetenzen im vorliegenden Beiheft ausmachen, sollen hinsichtlich der folgenden übergreifenden Fragen betrachtet werden:

- *Konzeptualisierung und Begründungszusammenhänge:* Was macht den inhaltlichen Kern des jeweiligen Konstrukts aus? Was ist der jeweilige Begründungszusammenhang, was macht das Konstrukt relevant?
- *Bezogen auf das methodische Vorgehen und die empirische Prüfung:* Wie wird die empirische Prüfung des Konstrukts (Skalierung) vorgenommen? Wie wird die Validitätsprüfung umgesetzt und wie wird Validität durch die erzielten Ergebnisse belegt? Welche Erkenntnisse liefern die Analysen? Welche weiteren Prüfungen und weiteren Forschungsarbeiten schließen sich an?

2.1 Konzeptualisierung und Begründungszusammenhänge

In den Blick genommen werden vier Konstrukte:

- *Argumentieren mit Evidenz im Bildungsbereich* (Teilkompetenzen Informationsauswahl und Bewertung von Studien) (Trempler, Hetmanek et al.),
- *bildungswissenschaftliche Forschungskompetenz* (Schladitz et al.),

- *Genrewissen über bildungswissenschaftliche Forschungspublikationen* (Winter-Hölzl et al.) und
- *Kompetenzen zum selbstregulierten Lernen im Studium* (Steuer et al.).

Beitrag 1: Vor dem Hintergrund der von verschiedenen Seiten geäußerten Forderung, auch im Bildungsbereich eine evidenzbasierte Praxis ähnlich wie in medizinischen Berufen zu etablieren (S. 145), gehen Trempler, Hetmanek et al. von der Anforderung aus, die sich für pädagogisches Personal stellt und von diesem erfolgreich bewältigt werden muss: „relevante wissenschaftliche Forschungsarbeiten zu finden, auszuwählen sowie mit Blick auf eine konkret anstehende Entscheidung zu bewerten und nutzbar zu machen" (S. 144). Für die dahinterstehende Kompetenz verwenden Trempler, Hetmanek et al. den Begriff des „evidenzbasierten Argumentierens" (S. 148); in ihrer Studie selbst werden davon, aufgrund ihrer grundlegenden Bedeutung, die beiden Teilkompetenzen „Informationsauswahl" und „Bewertung von Studien" fokussiert (S. 145). Als Abstraktionsebene wird die Auseinandersetzung mit empirischer Primärliteratur, d. h. empirische Forschungsartikel, gewählt, da sie als „primäre Quelle für forschungsbasierte Erkenntnisse" betrachtet werden (S. 146).

Beitrag 2: Auch in der Studie von Schladitz et al. spielt, ähnlich wie bei Trempler, Hetmanek et al., die Forderung nach evidenzbasierter Entscheidungsfindung in pädagogischen Kontexten als Begründungszusammenhang eine wichtige Rolle. In den Blick genommen wird die „bildungswissenschaftliche Forschungskompetenz" (BFK) als Konstrukt, wobei die Messung über drei Kompetenzfacetten erfolgt (S. 168–169 und Abb. 1): „Informationskompetenz" (z. B. kompetenter Umgang bei einer Schlagwortsuche), „statistische Kompetenz" (z. B. Lesen eines Balkendiagramms) und „auf Forschung bezogenes Kritisches Denken" (z. B. Analyse von schriftlichen, zusammenfassenden Darstellungen von Forschungsergebnissen).

Beitrag 3: Winter-Hölzl et al. zielen mit ihrer Studie auf die Entwicklung eines Tests zur Erfassung von Wissen über das Genre „Empirischer Forschungsartikel" bei Studierenden und Promovierenden der Bildungswissenschaften (S. 185). Sie verorten das zu modellierende Konstrukt, kognitionspsychologischen Modellen folgend und unter Bezug auf Modelle der Schreibforschung (S. 187), in Prozesse des wissenschaftlichen Schreibens. Die Autorinnen und Autoren weisen jedoch darauf hin, dass das Vorhandensein von Genrewissen noch kein Garant für gelingende Textproduktion sei (S. 199). Drei Wissensformen werden modellhaft unterschieden (Tab. 1) – deklaratives rhetorisches Wissen, diagnostisches rhetorisches Wissen, rhetorisches Wissen in der Textproduktion – und in einer Matrix mit inhaltlichen Wissensanforderungen gekreuzt; empirisch geprüft werden anschließend nur die ersten beiden Wissensformen.

Beitrag 4: Steuer et al. nehmen die Anforderung an Studierende, ein hohes Maß an eigenständiger Aneignung und Erarbeitung von Wissensbeständen zu zeigen, was charakteristisch für den Kompetenzerwerb im tertiären Bildungsbereich ist, als zentralen Ausgangspunkt für die Untersuchung von Kompetenzen zum selbstregulierten Lernen (SRL). Formuliert wird ein dreidimensionales Arbeitsmodell (Abb. 1), mit dessen Dimensionen „sich die Kompetenzen von Studierenden zum SRL (…) hinreichend um-

fassend beschreiben lassen“ (S. 207); indem die drei Dimensionen verknüpft werden, werde damit über den rein behavioralen Ansatz bisheriger Forschung hinausgegangen.

Während die ersten drei Beiträge die Gemeinsamkeit aufweisen, die theoretische Konzeptualisierung sowie Operationalisierung eines Kompetenztests einschließlich erster Analysen zur Validierung zum Gegenstand zu machen, setzt der vierte Beitrag von Steuer et al. mit dem Fokus auf die Formulierung eines mehrdimensionalen Arbeitsmodells an einer Problemstellung an, die erst zu einem nachfolgenden Zeitpunkt in die Entwicklung eines Kompetenztests münden soll.

Eine andere Variation zeigt sich bei den Beiträgen dahingehend, ob der jeweilige Begründungsrahmen für die fokussierte Anforderung, auf welche sich die zu modellierende Kompetenz bezieht, außerhalb der Hochschule im berufsbezogenen Kontext liegt oder ob es sich um Kompetenzen handelt, die für das Studium immanent sind und sich schon dadurch legitimieren lassen. Hat das „evidenzbasierte Argumentieren“ zum Beispiel letztlich zum Ziel, dass praktisch-professionell tätige Pädagogen wissenschaftliche Erkenntnisse in ihr Handeln einbeziehen, so ist die damit implizit unterliegende Wirkungskette vom Erwerb bis zur Anwendung der Kompetenz länger (und möglicherweise brüchiger) als zum Beispiel der Erwerb von Genrewissen über empirische Forschungsartikel und seine gezielte Nutzung während der Erstellung einer wissenschaftlichen Hausarbeit im Master-Studium. Hier stellt sich zwangsläufig die Frage, wie gesichert eine solche Wirkungskette theoretisch erarbeitet und empirisch untersucht werden kann.

Zentrales Ziel von TEDS-M war eine Bilanzierung der am Ende der Lehrerausbildung erreichten Ergebnisse, gleichwohl der Kompetenzmessung die Annahme unterlag, kognitive Dispositionen der Lehrkräfte zu testen, die als Voraussetzung für berufliches Handeln gelten, etwa zur Sicherstellung qualitativ hochwertiger Lerngelegenheiten (König & Pflanzl, eingereicht). In dem Follow-up zu TEDS-M (Blömeke, Kaiser & König, 2009), mit dem die in TEDS-M getesteten angehenden Lehrkräfte nach erfolgtem Berufseinstieg erneut getestet wurden, zeigte sich in längsschnittlichen Analysen, dass es zu erheblichen Umstrukturierungen und Transformationsprozessen der Lehrerkompetenzen während des Berufseinstiegs kommt (Blömeke et al., 2014; König et al., 2014). Die lange vermutete, empirisch jedoch nicht vollständig geprüfte Wirkungskette *Lehrerausbildung – Lehrerhandeln – Schülerleistung* (Blömeke, 2003; Terhart, 2012) erweist sich somit als komplex und brüchig. Für andere Berufsgruppen und Kompetenzbereiche sind Analogien vorstellbar (z. B. Bruer, 1997).

2.2 Methodisches Vorgehen – empirische Prüfung der Konstrukte

Eine spezifische Herausforderung aller vier Arbeiten ist die Operationalisierung der theoretischen Modellierung bzw. Konzeptualisierung des Konstrukts in Form des konkreten Erhebungsinstruments. Aus methodischen Gründen werden gelegentlich nur Ausschnitte des eigentlichen Konstrukts in den Blick genommen (z. B. Verzicht auf rhetorisches Wissen in der Textproduktion bei Winter-Hölzl et al.), und die Beiträge unter-

scheiden sich darin, wie detailliert über die psychometrische Prüfung des Instruments berichtet wird (besonders ausführlich: Winter-Hölzl et al.).

Items

Form, Inhalt und Aufbau von Testitems, ihre Zusammenstellung im Gesamtgefüge eines Testinstruments sind stets von fundamentaler Bedeutung. Verschiedene Entscheidungen der Testentwicklerinnen und -entwickler wie jene für ein bestimmtes Testitemformat (z. B. nur geschlossene Items) oder die Qualität der Distraktoren bei geschlossenen Items (z. B. single-choice vs. Rating-Skalen) können weitreichende Konsequenzen haben – auch für die späteren psychometrischen Eigenschaften des Testinstruments sowie die mit dem Instrument zu erzielenden Ergebnisse. Bei den hier in Rede stehenden Untersuchungen wurde unterschiedlich vorgegangen. Während Trempler, Hetmanek et al. nur geschlossene Items im Rating-Skalen-Format verwenden, nutzen Schladitz et al. Items im Multiple-Choice- und Fill-in-Format. Bei Winter-Hölzl et al. überwiegt der Anteil geschlossener Items gegenüber offenen Items, in der Analyse scheinen sich aber eher die offenen Items zu bewähren (10 von 14, ca. 70 %) als die geschlossenen (14 von 23, ca. 60 %).

In allen Beiträgen sind große Bemühungen erkennbar, unterschiedliche kognitive Anforderungen zu konzeptualisieren, die bei der Bearbeitung der Testitems durch die Probandinnen und Probanden eine Rolle spielen. Inwieweit mit der alleinigen Verwendung von geschlossenen Testitems komplexe kognitive Bearbeitungsprozesse erfasst werden können wie z. B. die Generierung von gedanklichen Alternativen zu einer Problemstellung, welche stellvertretend für tatsächliches Handeln in einer typischen Situation ist, bleibt eine große Herausforderung der Forschung zu wissenschaftsbezogenen Kompetenzen.

Beispielhaft wird dies sichtbar bei der Anforderung, Forschungsbefunde kritisch zu reflektieren als Teil bildungswissenschaftlicher Forschungskompetenz von Schladitz et al. Die Konzeptualisierung bezieht sich, dem Beispielitem folgend, vermutlich primär auf das genaue Verstehen des konkret Dargestellten. Diese Anforderung ist zweifellos ein wichtiger Bestandteil des Konstrukts. Von Interesse aber wäre weiterführend zum Beispiel, wie das kritische Reflektieren von Forschungsbefunden unter Hinzuziehen von testsituationsexternem Wissen (z. B. über wissenschaftliche Theorien oder Forschungsbefunde aus anderen Studien) getestet werden könnte. Die Ableitung von Hypothesen aus einem vorgegebenen Forschungsstand könnte möglicherweise ebenfalls die Verwendung von offenen Testformaten nahelegen, etwa wenn dies als kreativer Akt bildungswissenschaftlicher Forschungskompetenz betrachtet wird.

Auch TEDS-M stand vor der Herausforderung, nicht nur deklaratives Wissen zu testen, sondern auch den Versuch zu unternehmen, darüber hinaus wenigstens anteilig handlungsnahes Lehrerwissen zu erfassen. Hierfür eigneten sich komplexe, i. d. R. offene Testfragen, die es erfordern, auf eine kurze Schilderung einer typischen Problemsituation im Unterricht hin im offenen Antwortformat Handlungsstrategien oder -optionen zu nennen bzw. zu entwickeln (*generate* bzw. „Kreieren“). Diese Dimension kognitiver Anforderungen ließ sich über Subskalen der verwendeten Tests jeweils relia-

bel messen (Blömeke et al., 2010a, 2010b), und es lässt sich empirisch belegen, dass die Testleistungen an das Vorhandensein von Unterrichtserfahrung gekoppelt sind (König, 2013). TEDS-FU ging sogar noch einen Schritt weiter, indem die Kompetenzmessung durch videobasierte Instrumente erweitert wurde (Blömeke et al., 2014; König et al., 2014) und damit im Sinne des PID-Modells proximal Unterrichtsperformanz vorhergesagt werden kann (Blömeke, Gustafsson & Shavelson, im Druck).

Strukturanalysen zu den Konstrukten

Ergebnisinformationen aus strukturellen Analysen zur Binnendifferenzierung der Konstrukte sind in den Beiträgen von Trempler, Hetmanek et al. und Schladitz et al. vorhanden und aufschlussreich. Die Teilkompetenz „Informationsauswahl" wird über Testaufgaben gemessen, bei denen die Probandinnen und Probanden ihre Einschätzung bezüglich Ergebnissen von thematisch fokussierten Literaturrecherchen auf Likert-Skalen lokalisieren sollen. Die zweite Teilkompetenz „Bewertung von Studien" wird über Testaufgaben abgebildet, bei denen die Probanden Kurzfassungen empirischer Forschungsartikel anhand von Qualitätsaspekten bewerten sollen (ebenfalls über Likert-Skalen). Obwohl die eher niedrigen Skalen-Reliabilitäten (Cronbach's Alpha .65/.53) Anlass zur Vermutung geben, dass die „wahre" Korrelation etwas höher liegen dürfte, ist die manifeste Korrelation der beiden Teilkompetenzen mit .22 niedrig und legt eine analytische Trennung der beiden Teilkompetenzen nahe. Die Frage, inwieweit hierbei die unterschiedlichen Antwortformate zur Erfassung der beiden Teilkompetenzen *(mode-effect)* sowie die unterschiedlichen Scoring-Strategien eine Rolle spielen, könnte den Ausgangspunkt für eine interessante flankierende Skalierungsanalyse bilden.

Im Kontrast hierzu verweisen die bei Schladitz et al. sehr hohen Interkorrelationen der drei Teilkompetenzen (messfehlerbereinigt .80 und höher) sowie bei den Validierungsanalysen „kaum differenzierte Zusammenhänge zwischen Intelligenz (…) sowie den drei Kompetenzfacetten" (S. 179) auf Homogenität des erfassten Konstrukts. Da beide Studien, jene von Trempler, Hetmanek et al. und jene von Schladitz et al., thematisch Überschneidungen aufweisen, stellt sich die Frage nach binnendifferenzierenden Analysen weiterhin, etwa an anderen Stichproben. Möglicherweise wäre sogar eine gegenseitige Konstruktvalidierung aufschlussreich, wie sie von der Form her derzeit z. B. für das bildungswissenschaftliche und pädagogische Wissen angehender Lehrkräfte vorliegt (Seifert & König, 2012).

Binnendifferenzierende Strukturanalysen zu den erfassten Kompetenzkonstrukten sind auch in TEDS-M sowie in weiteren Studien zur professionellen Kompetenz von Lehrkräften wie der COACTIV-Studie (Kunter et al., 2011) von Bedeutung, wobei sie auf den multidimensionalen Charakter der kognitiven Voraussetzungen von Lehrkräften (Blömeke, 2014) und differenzielle Korrelationen zwischen den Dimensionen verweisen (Blömeke et al., 2010a, 2010b; Voss, Kunter & Baumert, 2011). In Analogie wäre auch für die hier fokussierten Kompetenzbereiche interessant, spezifische Annahmen zur Binnenstruktur der untersuchten Kompetenzen zu entwickeln und zu prüfen.

Stichprobe: Heterogenität und Homogenität von Studierenden

Korrelationsergebnisse sind auch abhängig von der verwendeten Stichprobe, z. B. der Heterogenität ihrer Probanden in Bezug auf die getesteten Kompetenzen. Es macht einen Unterschied, nur Bachelor-Studierende des 5. Semesters mit dem Studienschwerpunkt Bildungswissenschaften einzubeziehen oder aber, wie es z. B. bei Trempler, Hetmanek et al. der Fall ist, darüber hinaus Studierende in unterschiedlichen Ausbildungsstadien (BA, MA, Promotionsstudium). Insbesondere wenn curriculare Validität der getesteten Kompetenz angenommen wird, kann die Kontrolle des Ausbildungsstadiums von Relevanz sein. Die unerwartet niedrige Korrelation zwischen Selbsteinschätzung und objektiver Testung bei Schladitz et al. wäre also vor dem Hintergrund der Stichprobenzusammensetzung zu diskutieren. Studierende im 3. Semester (Median), von denen über 60 % ein Lehramt anstreben, wurden noch nicht sonderlich umfangreich mit bildungswissenschaftlichen Forschungsmethoden konfrontiert. Angesichts eher einführender Veranstaltungen zu Beginn der Ausbildung (Darge, Schreiber, König & Seifert, 2012) werden vermutlich nur wenige Lehramtsstudierende einer Aussage wie „Ich fühle mich sicher in der Formulierung wissenschaftlicher Fragestellungen und Hypothesen" voll zustimmen können. Da eine verlässliche Selbsteinschätzung unter anderem auf Erfahrungen beruht, könnte die Erfassung selbst durch andere Faktoren (z. B. Selbstwirksamkeit) geprägt sein (vgl. z. B. Cramer, 2010). Auch in TEDS-M (König, Kaiser & Felbrich, 2012) und in der Folgestudie LEK (König & Seifert, 2012) für die erste Phase der Lehrerausbildung konnten nur schwache Zusammenhänge zwischen Wissen und Selbsteinschätzungen gefunden werden, welche die alleinige Erfassung von Kompetenzen durch Selbstberichte infrage stellen.

Externe Kriterien zur Validierung der Tests

Trotz vielfacher Kritik am Kompetenzbegriff ist es ein wertvolles Verdienst, mit ihm eine konzeptionelle Abgrenzung zum Intelligenzbegriff vorliegen zu haben (Weinert, 2001; Hartig & Klieme, 2006), vor allem weil die empirische Bildungsforschung an der Erfassung von Leistungen interessiert ist, die als Resultate von Lehr-Lernprozessen betrachtet werden können und demnach auch beeinflussbar sind. Demnach sind Analysen zur diskriminanten Validität der entwickelten Kompetenzkonstrukte mit Blick auf allgemeine kognitive Fähigkeiten von Bedeutung. In dieser Hinsicht kommen die Studien zu nachvollziehbaren und plausiblen Ergebnissen.

Hinsichtlich der weiteren externen Kriterien zur Validierung der Tests wäre es möglicherweise interessant, vor allem auf bereits vorhandene und etablierte Instrumente (z. B. zur Erfassung des pädagogischen Wissens bei Trempler, Hetmanek et al. mithilfe des TEDS-M Instruments, König & Blömeke, 2010) bei der Validierung mithilfe von Außenkriterien zurückzugreifen, da diese bereits bekannte Messeigenschaften aufweisen und somit die eigenen Validierungsanalysen aussagekräftiger in den Forschungsdiskurs eingeordnet werden können. Über die dargestellten Analysen zur diskriminanten und konvergenten Validität hinaus stellt sich die Frage nach Belegen zur prognostischen Validität.

3. Weiterführende Fragen

Die vier Beiträge unterstreichen die Wichtigkeit, wissenschaftsbezogene Kompetenzen zu modellieren und zu erfassen. Sie bilden unentbehrliche Grundlagen, auf denen wichtige Fragen zur Qualitätsentwicklung und -sicherung im Hochschulbereich erst bearbeitet werden können. Im Sinne vorhandener Initiativen zur bilanzierenden Untersuchung von tertiärer Bildung wie TEDS-M und den Folgestudien im Kontext von TEDS-M wäre es weiterführend von Interesse, folgende Fragen zu berücksichtigen:

- Wie können Erwartungen, die mit den Konzeptionen der entwickelten Instrumente an eine „Wirkungskette" von (Erst-)Erwerb bis professioneller Anwendung der jeweiligen Kompetenz verbunden sind, theoretisch formuliert und empirisch analysiert werden?
- Wie können die Instrumente weiterentwickelt oder ergänzt werden, um auch (besonders) komplexe kognitive Kompetenzen messen zu können?
- Welche binnendifferenzierenden Strukturen sind den fokussierten Konstrukten eigen und wie können diese empirisch abgebildet werden?
- Auf welche Zielgruppe sind die Kompetenzen zugeschnitten und welche Stichprobendefinition muss gewählt werden, um belastbare deskriptive Aussagen (z. B. von nach einem bestimmten Bildungsabschnitt erreichten Kompetenzen) treffen zu können?
- Welche prognostische Validität der Instrumente wird angenommen und wie können hierzu Belege erbracht werden?

Die Projekte sind von hoher Qualität und besitzen zweifellos ein großes Potenzial, sodass ein erheblicher Fortschritt in der Konzeptualisierung und empirischen Untersuchung wissenschaftsbezogener Kompetenzen in den kommenden Jahren auch zu diesen weiterführenden Fragen erwartet werden darf.

Literatur

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

Blömeke, S. (2003). *Lehrerausbildung – Lehrerhandeln – Schülerleistungen. Perspektiven nationaler und internationaler empirischer Bildungsforschung* (Antrittsvorlesung). Berlin: Humboldt-Universität. http://edoc.hu-berlin.de/humboldt-vl/139/bloemeke-sigrid-3/PDF/bloemeke.pdf [14. 11. 2014].

Blömeke, S. (2014). *Modelling teachers' professional competence as a multi-dimensional construct* (paper presented at the symposium on „Teachers as Learning Specialists – Implications for Teachers' Pedagogical Knowledge and Professionalism" hosted by OECD's Centre for Educational Research and Innovation and the Flemish Department of Education and Training. Brussels, Belgium, 18 June 2014).

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (im Druck). Beyond dichotomies: Viewing competence as a continuum. *Zeitschrift für Psychologie*.

Blömeke, S., Kaiser, G., & König, J. (2009). *Längsschnittliche Entwicklung der Kompetenzen von Junglehrkräften: Follow-Up zur internationalen Vergleichsstudie TEDS-M* (TEDS-FU). Antrag auf Gewährung einer DFG-Sachbeihilfe im Rahmen des Normalverfahrens (Ms.; bewilligt als BL 548/8e1). Berlin/Hamburg/Köln: Universität.

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

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

Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M., & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf – Fachbezogenes Wissen als Voraussetzung für Wahrnehmung, Interpretation und Handeln im Unterricht. *Zeitschrift für Erziehungswissenschaft, 17*(3), 509–542.

Bruer, J. T. (1997). Education and the brain: A bridge too far. *Educational Research, 26*, 4–16.

Cramer, C. (2010). Kompetenzerwartungen Lehramtsstudierender. Grenzen und Perspektiven selbsteingeschätzter Kompetenzen in der Lehrerbildungsforschung. In A. Gehrmann, U. Hericks & M. Lüders (Hrsg.), *Bildungsstandards und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 85–97). Bad Heilbrunn: Klinkhardt.

Darge, K., Schreiber, M., König, J., & Seifert, A. (2012). Lerngelegenheiten im erziehungswissenschaftlichen Studium. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* (S. 87–118). Münster: Waxmann.

Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 128–143). Heidelberg: Springer.

Klieme, E. (2012). Internationales large scale assessment in der Lehrerbildung: Anmerkungen zu einem neuen Paradigma der vergleichenden Bildungsforschung. *Zeitschrift für Pädagogik, 58*(4), 492–499.

König, J. (2013). First comes the theory, then the practice? On the acquisition of general pedagogical knowledge during initial teacher education. *International Journal of Science and Mathematics Education, 11*(4), 999–1028.

König, J., & Blömeke, S. (2010). *Pädagogisches Unterrichtswissen (PUW). Dokumentation der Kurzfassung des TEDS-M-Testinstruments zur Kompetenzmessung in der ersten Phase der Lehrerausbildung*. Berlin: Humboldt-Universität.

König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A., & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education, 38*, 76–88.

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

König, J., & Pflanzl, B. (eingereicht). *Is teacher knowledge associated with teacher performance? On the relationship between teachers' general pedagogical knowledge and instructional quality.*

König, J., & Seifert, A. (Hrsg.) (2012). *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung.* Münster: Waxmann.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Seifert, A., & König, J. (2012). Pädagogisches Unterrichtswissen – bildungswissenschaftliches Wissen: Validierung zweier Konstrukte. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* (S. 215–233). Münster: Waxmann.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Research, 57*, 1–22.

Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., &Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries. Findings from the IEA teacher education and development study in mathematics (TEDS-M)*. http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/TEDS-M_International_Report.pdf [14. 11. 2014].

Terhart, E. (2012). Wie wirkt Lehrerbildung? Forschungsprobleme und Gestaltungsfragen. *Zeitschrift für Bildungsforschung, 2*(1), 3–21.

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). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessung in Schulen* (S. 17–31). Weinheim/Basel: Beltz.

Wissenschaftsrat (2000). *Empfehlungen zur Einführung neuer Studienstrukturen und -abschlüsse (Bakkalaureus/Bachelor – Magister/Master) in Deutschland*. Berlin: Wissenschaftsrat.

Anschrift des Autors

Prof. Dr. Johannes König, Universität zu Köln, Humanwissenschaftliche
Fakultät, Institut für Allgemeine Didaktik und Schulforschung,
Gronewaldstraße 2, 50931 Köln, Deutschland
E-Mail: johannes.koenig@uni-koeln.de

Leistungsbeurteilung in Theorie und Praxis

Eiko Jürgens/Urban Lissmann
Pädagogische Diagnostik
Grundlagen und Methoden der Leistungsbeurteilung in Schulen
2015. 208 Seiten. Broschiert.
ISBN 978-3-407-25708-6

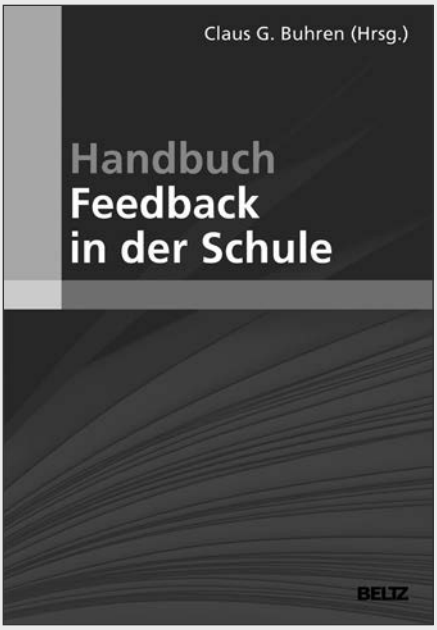
Die reine Ziffernote als Ergebnis der Leistungsbeurteilung wird schon lange nicht mehr den Anforderungen gerecht, die an einen modernen kompetenz- und heterogenitätsorientierten Unterricht gestellt werden. In einem Unterricht, der auf »Lernen lernen« ausgerichtet ist, müssen die Leistungen, Fähigkeiten und Fertigkeiten differenziert ermittelt und bewertet werden. Hierzu liefert das Buch von Eiko Jürgens und Urban Lissmann einen kompakten und stets praxisbezogenen Überblick. Es stellt anschaulich klassische und alternative Verfahren und Instrumente der Leistungsbeurteilung dar und setzt sie in Bezug zu den Standards der Lehrerausbildung.

Aus dem Inhalt
- Die Beurteilungsaufgabe im Lehrerberuf
- Grundlagen der pädagogischen Diagnostik
- Methoden der Leistungserfassung und Leistungskontrolle
- Prüfungen vorbereiten, durchführen und analysieren
- Beurteilungsformen

BELTZ

Beltz Verlag · Weinheim und Basel · Weitere Infos und Ladenpreis: www.beltz.de

Alles, was Sie über Feedback wissen müssen



Claus G. Buhren (Hrsg.)
Handbuch Feedback in der Schule
2015. 480 Seiten. Gebunden.
ISBN 978-3-407-83186-6

Die Hattie-Studie hat bestätigt, dass Feedback in der Schule eine besonders wirkungsvolle Maßnahme ist, um die Unterrichtsqualität und den Lernerfolg der Schülerinnen und Schüler zu steigern. PISA-Sieger wie Schweden, Finnland, Kanada oder Japan verdanken ihren Erfolg nicht zuletzt einer ausgeprägten Feedbackkultur in der Schule.

Dieses Handbuch beschreibt die unterschiedlichen Formen und Methoden von Feedback (Führungs-, Schüler-, Lehrer- und Elternfeedback) und ergänzt sie durch konkrete Beispiele aus der Praxis. Die Autorinnen und Autoren erläutern den Begriff des Feedbacks, seine Entstehung und damit verbundene Konzepte von Schul- und Unterrichtsentwicklung. Aktuelle Forschungsergebnisse, Schritte zum Aufbau einer Feedbackkultur in der Schule und Beispiele aus der Schulpraxis runden den Band ab.

Dr. Claus G. Buhren ist Professor für Schulentwicklung an der Deutschen Sporthochschule Köln und wissenschaftlicher Leiter der Deutschen Akademie für Pädagogische Führungskräfte (DAPF) der TU Dortmund.

BELTZ

Beltz Verlag · Weinheim und Basel · Weitere Infos und Ladenpreis: www.beltz.de

Die aktuelle und neue Einführung in die Statistik

Mit Online-Materialien
Grundlagentexte Soziologie
2014, 272 Seiten
broschiert
€ 19,95
ISBN 978-3-7799-2613-9

Diese Einführung in die Statistik orientiert sich an der Praxis sozialwissenschaftlicher Datenanalyse. Die mathematischen Grundlagen werden soweit dargestellt und erläutert, wie es zum Verständnis der Verfahren notwendig ist. Im Vordergrund steht aber der Umgang mit Daten. Stets geht es um die Fragen: Welche Bedeutung haben die Verfahren, Kennzahlen und Graphiken der univariaten, bivariaten und multivariaten Datenauswertung? Wie kann man Aussagen statistisch absichern? Da statistische Datenauswertung heute ohne einschlägige Software nicht mehr denkbar ist, erläutert das Buch auch die Umsetzung der Verfahren mit zwei bekannten Softwarepaketen, SPSS und Stata.

Aus dem Inhalt:

Daten, Forschungsdesigns und Stichproben
Univariate Analyse
Von der Stichprobe zur Grundgesamtheit: Statistisches Schließen
Bivariate Analyse
Regressionsanalyse

www.beltz.de

BELTZ JUVENTA