Jenßen, Lars; Dunekacke, Simone; Blömeke, Sigrid

Qualitätssicherung in der Kompetenzforschung. Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis

Blömeke, Sigrid [Hrsg.]; Zlatkin-Troitschanskaia, Olga [Hrsg.]: Kompetenzen von Studierenden. Weinheim u.a. : Beltz Juventa 2015, S. 11-31. - (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

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.

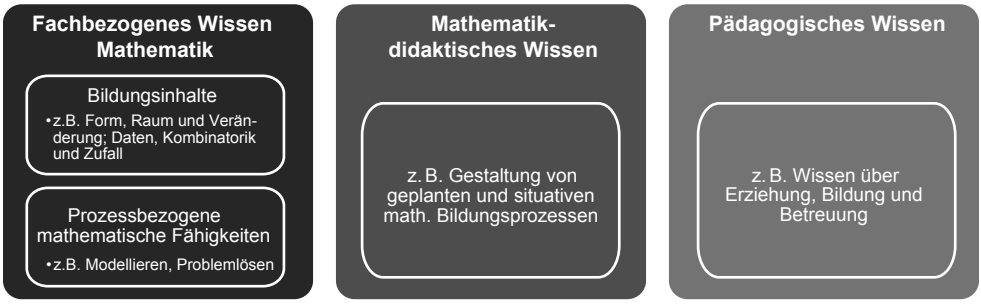

Abb. 1: Kurzfassung der kognitiven Facetten des Kompetenzstrukturmodells in KomMa

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).

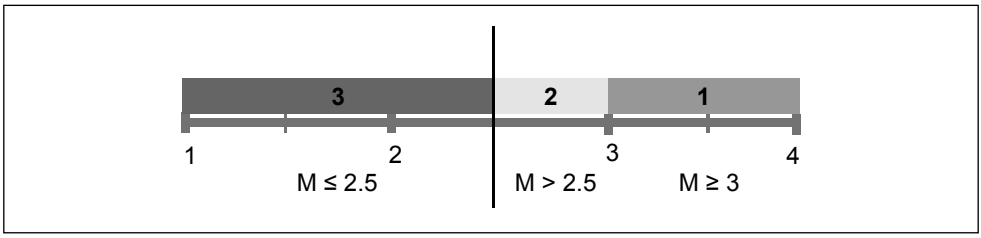

Abb. 2: Auswertungsschema zur Inhaltsvalidität in KomMa

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