

61. Beiheft

April 2015

ZEITSCHRIFT FÜR PÄDAGOGIK

Kompetenzen
von Studierenden

BELTZ JUVENTA

Leseprobe aus: Zeitschrift für Pädagogik, 61. Beiheft, ISBN 978-3-7799-3507-0

© 2015 Beltz Verlag, Weinheim Basel

<http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-3507-0>

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-

8 Blömeke/Zlatkin-Troitschanskaia: Einleitung

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 mehrbenenanalytische 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-

10 Blömeke/Zlatkin-Troitschanskaia: Einleitung

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, Duneck 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-

12 Jenßen/Dunekacke/Blömeke: Qualitätssicherung in der Kompetenzforschung

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 zu einem 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 Testverfahren¹

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 Testwertinterpretation in den Vordergrund. Validierung wird demnach definiert als „the appropriateness, 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 einzu beziehen. 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-

14 Jenßen/Dunekacke/Blömeke: Qualitätssicherung in der Kompetenzforschung

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 *Testentwicklung* 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

16 Jenßen/Dunekacke/Blömeke: Qualitätssicherung in der Kompetenzforschung

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 & Himelarb, 2011).

Zeitlich gesehen kann bei einer kriterialen Validierung zwischen konkurrender und prognostischer Validität unterschieden werden (Schaper, 2013). Im Falle der Feststellung von konkurrender 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).