KoKoHs

Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor

www.kompetenzen-im-hochschulsektor.de

KoKoHs Working Papers

Nr. 2

Sigrid Blömeke

Validierung als Aufgabe im Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Johannes Gutenberg-Universität Mainz

Humboldt-Universität zu Berlin

KoKoHS Working Papers zur BMBF-Förderinitiative
„Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“

In den KoKoHs Working Papers werden Beiträge veröffentlicht, welche in Zusammenhang mit der Förderinitiative „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“ entstanden sind. Es kann sich dabei u.a. um konzeptionelle Arbeiten und erste Befunde handeln, welche im Interesse einer schnellen Verbreitung angezeigt werden oder öffentlich diskutiert werden sollen. Die Veröffentlichungen als Working Paper schließen nicht aus, dass die Texte anderweitig veröffentlicht werden. Die Verantwortung für die Inhalte liegt in der Hand der Autorinnen und Autoren. Sie geben nicht notwendigerweise die Position der Koordinierungsstelle KoKoHs als Herausgeber wider.

Herausgeberinnen:

Prof. Dr. Sigrid Blömeke
Humboldt Universität zu Berlin
Philosophische Fakultät IV
Abt. Systematische Didaktik und Unterrichtsforschung
Unter den Linden 6
D-10099 Berlin

Prof. Dr. Olga Zlatkin-Troitschanskaia
Johannes Gutenberg Universität Mainz
Fachbereich 03: Rechts- und Wirtschaftswissenschaften
Lehrstuhl für Wirtschaftspädagogik I
Jakob Welder-Weg 9
D-55099 Mainz

Kontakt:

susanne.schmidt@uni-mainz.de,
corinna.lautenbach@hu-berlin.de

Das Vorhaben wird aus Mitteln des Bundesministeriums für Bildung und Forschung unter den Förderkennzeichen 01PK11100A und 01PK11100B gefördert.

Die Working Papers stehen auch als Download zur Verfügung:

http://www.kompetenzen-im-hochschulsektor .de

Validierung als Aufgabe im Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“

Sigrid Blömeke

Kontakt:

susanne.schmidt@uni-mainz.de,
corinna.lautenbach@hu-berlin.de

Bibliographische Angaben:

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

Validierung als Aufgabe im Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“

Zusammenfassung:
Validität ist das fundamentalste und zugleich komplexeste Gütekriterium empirischer Bildungsforschung. Bereits die klassische Definition von Validität als das Ausmaß, in dem ein Testverfahren misst, was es messen soll, war vergleichsweise schwierig zu überprüfen. Die Komplexität der Validierung von Messverfahren ist in den letzten Jahren durch das Wiederaufgreifen früher geführter Kontroversen nochmals gesteigert worden. Das vorliegende Working Paper legt den Diskussionsstand bezogen auf den Bereich der Bildungsforschung und den Bereich der Leistungsdiagnostik in der Hochschulforschung dar. Diese Zusammenfassung beruht auf den Vorträgen und Workshops im Rahmen des Rundgesprächs der Förderinitiative „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“ an der Humboldt-Universität zu Berlin im März 2013. Darüber hinaus erfolgt eine Auswertung zentraler Diskussionsbeiträge in den einschlägigen psychometrischen Standardwerken und Zeitschriften.

Schlagworte:
Validierung, Validität, Kompetenzerfassung , tertiärer Bildungssektor

Abstract:
Validity as the most basic criteria of quality in empirical educational research is at the same time the most complex one. Traditionally, validity defines the extent to which a test measures what it is supposed to, which is already relatively difficult to prove. In the last few years validating test procedures has become even more complex as past controversies have been reactivated. The present working paper summarizes the current state of the debate on validity in the fields of educational research and cognitive ability testing. It is based on lectures and workshops at the Roundtable Conference of the research projects included in the funding initiative “Modeling and Measuring Competencies in Higher Education” at Humboldt University in Berlin in March 2013. In addition, essential contributions to the current discussion in relevant psychometric journals are analyzed.

Keywords:
Validation, Validity, competence assessment, tertiary education

.

1 Vom klassischen Validitätsbegriff zum erweiterten Verständnis von Validität

Validität ist das fundamentalste und zugleich komplexeste Gütekriterium empirischer Bildungsforschung. Während für die Sicherung von Objektivität (Unabhängigkeit eines Testergebnisses von den Rahmenbedingungen der Testung) und Reliabilität (Zuverlässigkeit des Ergebnisses) als den Voraussetzungen valider Messungen weitgehend einheitlich definierte Begriffsverständnisse und Prüfverfahren vorliegen, stellt sich die Situation für den Prozess der Validierung anders dar. Bereits die klassische Definition von Validität als das Ausmaß, in dem ein Testverfahren misst, was es messen soll, war vergleichsweise schwierig zu überprüfen. Die Komplexität der Validierung von Messverfahren ist in den letzten Jahren durch das Wiederaufgreifen früher geführter Kontroversen zudem nochmals gesteigert worden.

Insbesondere Kane (2006, 2103) tritt im Anschluss an Cronbach (1971) und Messick (1989) dafür ein, die geplante *Verwendung* von Testscores stärker in den Fokus der Validierung zu rücken als das Messinstrument, mit dem dieser generiert worden ist. Er will damit darauf aufmerksam machen, dass die Interpretationen eines Scores und die daraus gezogenen Konsequenzen unterschiedlich und ggf. zu weitreichend sein können. Kane fordert daher: Je weitreichender die Schlussfolgerung aus einem Testergebnis sein soll, umso stärker muss der Nachweis erbracht werden, dass sie durch das Untersuchungsdesign tatsächlich gerechtfertigt sind. Die Standesvereinigungen American Educational Research Association, American Psychological Association und National Council on Measurement in Education haben diese Sichtweise bereits zum Teil in ihre Standards für die Verwendung von Testverfahren aufgenommen (AERA, APA & NCME, 1999) und werden sie in der für nächstes Jahr zu erwartenden Neuauflage voraussichtlich weiter stärken (www.teststandards.net/resources.htm).[1]

Die folgende Zusammenfassung des Diskussionsstands, die auf den Bereich der Bildungsforschung und weiter auf den Bereich der Leistungsdiagnostik in der Hochschulforschung zugeschnitten ist[2], beruht zum ersten auf dem Überblicksvortrag von Frey (2013), den dieser im Rahmen des Rundgesprächs der Förderinitiative „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor“ an der Humboldt-Universität zu Berlin gehalten hat. Zum zweiten stellen die Präsentationen in drei vertiefenden Workshops zur Konstrukt-, Kriteriums- und Inhaltsvalidierung (Hartig, 2013; Schaper, 2013; Schwippert, 2013) eine Basis dar. Und drittens erfolgte eine Auswertung zentraler Diskussionsbeiträge in den einschlägigen psychometrischen Standardwerken und Zeitschriften, insbesondere von Borsboom, Mellenbergh und Van Heerden (2004) sowie Scriven (2010) als Vertretern des eher traditionellen Konzepts der Instrumentvalidierung und Kane als dem Vertreter des Ansatzes der Interpretationsvalidierung.

Drei bis heute gebräuchliche – und auch von Kane nicht in Frage gestellte – Validitätskriterien weisen die längste Tradition in der Leistungsdiagnostik auf (Frey, 2013): die Kriteriumsvalidierung, die Inhaltsvalidierung und die Konstruktvalidierung von Testverfahren. Auf diese wird im Folgenden zunächst eingegangen, bevor neuere Entwicklungen nachgezeichnet und in ihrer Bedeutung für die Kompetenzerfassung im Hochschulsektor analysiert werden. Abschließend wird die aktuelle Kontroverse um den Validitätsbegriff zusammengefasst.

---

[1] Andere bedeutende Arbeiten zur Validierung, die herausgehoben werden sollen, stammen von Campbell und Fiske (1959), Anastasi (1986) sowie Cronbach (1989).

[2] Diese Zuspitzung erfolgt, da dem Begriff der Validität in anderen Kontexten andere Bedeutung zukommt. Dies gilt sowohl für andere Disziplinen, beispielsweise in der Philosophie oder in Jura, aber auch für andere Forschungsparadigmen, insbesondere die qualitative Forschung.

Kriteriumsvalidität

Mit Kriteriumsvalidität wird die Vorhersage einer direkt beobachtbaren Verhaltensweise außerhalb der Testsituation – daher auch als externe Validität bezeichnet – 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 von Vertragsabschlüssen), 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 das Außenkriterium erklären. Dabei werden Testwert und Kriteriumswert zum selben Zeitpunkt erhoben. 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 späteres Verhalten oder Leistungen vorhergesagt werden können. Das Kriterium wird dabei zu einem späteren Zeitpunkt erhoben als der Testwert. Ein Beispiel hierfür stellt die Vorhersage von Studienerfolg anhand eines Studieneingangstests dar. Von spezifischem Interesse ist dabei häufig, inwieweit das Testverfahren hilfreich ist, wenn es *zusätzlich* zu bekannten Maßen eingesetzt wird (inkrementelle Validität). Im Falle von Studieneingangstests würde dann beispielsweise gefragt, inwieweit ihnen inkrementelle Validität in Ergänzung zur Abiturnote zukommt, die vergleichsweise leicht erhoben werden kann und deren prognostische Validität für Studienerfolg vielfach belegt ist.

Liegt kriteriale Validität vor, können also nicht nur Aussagen über die Gültigkeit eines Testverfahrens und die Generalisierbarkeit der mit ihm gewonnenen Ergebnisse gemacht werden, sondern es wird auch möglich, Prognosen oder Diagnosen in Bezug auf das Verhalten oder die Leistungsfähigkeit in zukünftigen Kontexten und Anforderungsbereichen zu machen (Schaper, 2013). Der Grad, mit dem Kriterien aufgrund eines Testergebnisses prädiziert werden können, hängt dabei von der Objektivität und Reliabilität der Messung des Prädiktors sowie von der Objektivität, Reliabilität sowie Inhalts- und Konstruktvalidität des Kriteriums ab (Nerdinger, Blickle & Schaper, 2011).

Typische Probleme mit der Objektivität einer Kriteriumsmessung sind z.B. subjektive Urteilstendenzen (z.B. Halo-Effekte) von Fremdbeurteilungen. Ein typisches Problem mit der Reliabilität ist, dass sich das Kriterium auf Verhaltensmaße bezieht, die nicht stabil sind, sondern im Zeitverlauf oder situationsabhängig variieren (z.B. die Art und Qualität der Klassenführung bei Lehrkräften). Ein typisches Problem der Konstruktvalidität ist schließlich, dass Leistungskriterien wie Schulnoten in der Regel das Resultat eines Zusammenspiels individuellen Leistungsverhaltens mit weiteren (z.B. umgebungsbezogenen) Einflussfaktoren sind und keine Theorie zu deren Zusammenwirken vorliegt (alle Beispiele aus Schaper, 2013).

Mit diesem letzten Problem ist zudem die generellere Herausforderung angesprochen, die inhaltliche *Relevanz* eines Kriteriums zu sichern. Mit Kriteriumsrelevanz ist das Ausmaß gemeint, in dem das Kriterium einen wichtigen Aspekt des Konstrukts erfasst, wenn beispielsweise die Kundenzufriedenheitsbewertung als Kriterium für die Serviceorientierung von Call-Center-Agenten verwendet wird. Ist der Merkmalsbereich durch das Kriterium nicht vollständig abgedeckt, spricht man von Kriteriumsdefizienz. Ein Beispiel hierfür wäre, dass die angesprochene Kundenzufriedenheitsbewertung nicht erfasst, was der Call-Center-Agent vorbereitend oder im Anschluss an das Gespräch zur Erfüllung der Kundenwünsche tut. Spiegelt das Kriterium andere Merkmalsaspekte wieder als gemeint, indem Kundenzufriedenheitsbewertungen zu einem Call-Center-Agenten möglicherweise auch die Zufriedenheit des Kunden mit dem gesamten Unternehmen beinhalten, spricht man von Kriteriumskontamination (Schaper, 2013).

Speziell im Bereich der Kompetenzforschung, die im Zuge der Förderinitiative KoKoHs durchgeführt wird, stellt sich als zusätzliche Herausforderung, dass zwar angenommen wird, Kompetenz unterliege als latente Disposition erfolgreichem Handel in Realsituation (Performanz), dass zwischen diesen beiden Merkmalen aber zahlreiche vermittelnde Prozesse stattfinden, die einen direkten Nachweis schwierig machen. Wass et al. (2001) und Albino et al. (2008) haben daher – Millers Kompetenzpyramide weiterführend – einen vierschrittigen Validierungsprozess vorgeschlagen, der sequentiell Zwischenschritte wie kognitive Umstrukturierungen und unterschiedliche Rahmenbedingungen berücksichtigt und damit erfolgsversprechender zu sein scheint. Miller (1990) hat, aufbauend auf Blooms Taxonomie kognitiver Prozesse, dem Grad der Authentizität nach methodologisch zwischen vier Assessment-Formaten unterschieden, mit denen professionelle Kompetenz – in diesem Falle von Medizinern – untersucht werden kann:

1) Wissentests, in denen die angehenden Mediziner das Vorhandensein von deklarativem Wissen (*factual knowledge*) demonstrieren (*know*);
2) Kompetenztests im engen Sinne, in denen das Vorhandensein von anwendungsbezogenem Wissen (*applied knowlege*) demonstriert werden muss (*know how*);
3) Performanztests, in denen die angemessene Umsetzung des Wissens situiert in repräsentativen berufsbezogenen Situationen (*application of knowledge coupled with skills and the appropriate attitudes*) demonstriert werden muss (*show how*); und
4) handlungsbezogene Tests, in denen die angemessene Umsetzung des Wissens unter den Bedingungen des beruflichen Alltags demonstriert werden muss (*does*).

Diese vier Assessment-Formate können methodisch wie folgt durchgeführt werden (Wass et al., 2001; Albino et al., 2008):

1) In einem ersten Schritt wird kontextfrei das vorhandene, in der universitären Ausbildung erworbene (deklarative) Wissen erfasst (*factual recognition*). Dies kann mithilfe ökonomisch einsetzbarer Testverfahren, z.B. Multiple-Choice-Items, geschehen oder mithilfe schriftlicher Erhebung bzw. mündlicher Verfahren.
2) In einem zweiten Schritt geht es um die Untersuchung des Zusammenhangs zwischen deklarativem Wissen und dem stärker prozeduralisierten, anwendungsorientierten Wissen (*capacity for context application*). Dies kann über Testverfahren geschehen, die – noch immer standardisiert und in Form von Multiple-Choice-Items – die Wissensanwendung in typischen beruflichen Anforderungen situieren oder über das Verfassen freier Essays.

3) Einen weiteren Schritt näher an das tatsächlich zu erwartende Handeln kommen *performance assessments in vitro*, also Assessments unter kontrollierten Bedingungen. Hier sind Testteilnehmer gefordert, konkrete Lösungen für simulierte Alltagssituationen (zum Beispiel präsentiert in Form von Simulationen, Laborexperimenten oder über Videovignetten mit authentischen Szenen) zu generieren (als Beispiele im Lehrerkontext siehe Blömeke et al., eingereicht). Hier schließt der Lösungsprozess die Wahrnehmung und Analyse der Situation sowie die Reaktion auf diese ein.
4) Erst im letzten Schritt geht es schließlich um den Zusammenhang zum tatsächlichen Handeln im beruflichen Alltag (*performance assessment in vivo*). Hier geht es um eine holistische Betrachtung dessen, was gekonnt wird, beispielsweise über Videoaufnahmen dokumentiert und dann standardisiert analysiert (als Beispiel im Lehrerkontext siehe Vogelsang & Reinhold, 2012).

Inhaltsvalidität

Historisch gesehen fand neben der Kriteriumsvalidierung zunächst insbesondere die Sicherung von Inhaltsvalidität Anwendung (Frey, 2013). Hierbei geht es darum, dass Experten beurteilen, ob ein Test angemessen das interessierende Merkmal operationalisiert. In einer abgeschwächten, nicht strikt standardisiert durchgeführten Variante wird Inhaltsvalidität auch als Augenscheinvalidität (*face validity*) bzw. logische Validität bezeichnet (Schwippert, 2013).

Der Feststellung von Inhaltsvalidität kommt vor allem Bedeutung zu, wenn dem zu erfassenden Konstrukt mehr als nur eine operationale Definition zugrunde liegt (Schwippert, 2013). Bei letzterem Vorgehen ist die Natur des Konstrukts lediglich durch die Testitems definiert: der Test misst, was er misst; weiterreichende Ansprüche sind mit ihm nicht verbunden. Anders sieht dies im Falle theoretischer Begründungen eines Tests aus, bei einem hypothetisch-deduktiven Vorgehen also. Dann unterliegt dem Diagnoseverfahren eine Theorie, aus der konkrete Facetten und Dimensionen des zu erfassenden Konstrukts abgeleitet werden, die wiederum in Items umgesetzt (operationalisiert) werden. Unterschiede in der Testleistung sollen dann mithilfe der theoretischen Begründung erklärt werden können. Um dies zu leisten, muss allerdings belegt werden, dass der Test tatsächlich inhaltlich valide ist, dass die verwendeten Items also die konkreten Facetten und Dimensionen des Konstrukts und damit das Gesamtkonstrukt repräsentieren.

Zwei Probleme gefährden die Inhaltsvalidität eines Testverfahrens besonders häufig (Schwippert, 2013): das Konstrukt ist durch die Items nicht hinreichend im Test repräsentiert oder die Items führen Varianz in den Test ein, die konstrukt-irrelevant ist. Ein Beispiel für den ersten Fall wäre, wenn in einem Test zur Erfassung von fachbezogenen Lehrerkompetenzen das mathematikdidaktische Professionswissen außen vor gelassen und lediglich das mathematische Wissen erfasst würde. Ein Beispiel für den zweiten Fall wäre, wenn in dem Test zur Erfassung von fachbezogenen Lehrerkompetenzen zusätzliche Aspekte, wie beispielsweise Wissen zur Geschichte des 19. Jahrhunderts erfasst würden, die von der Konstruktdefinition nicht gedeckt sind (soziale Kompetenzen).

Wie bedeutsam die Feststellung von Inhaltsvalidität ist, macht ein populäres Beispiel deutlich (Schwippert, 2013). „PISA beansprucht, Basiskompetenzen zu erfassen, die in modernen Gesellschaften für eine befriedigende Lebensführung in persönlicher und wirtschaftlicher Hinsicht sowie für eine aktive Teilnahme am gesellschaftlichen Leben notwendig sind.“ (Deutsches PISA-Konsortium, 2000, S.

29). Zu diesem Zweck werden Basiskompetenzen im Lesen, in Mathematik und in Naturwissenschaften erfasst. Soweit bekannt, liegen Studien zum Zusammenhang dieser Kompetenzen mit dem Ausmaß an „befriedigende(r) Lebensführung in persönlicher und wirtschaftlicher Hinsicht sowie für eine aktive Teilnahme am gesellschaftlichen Leben“ noch nicht vor.

In der Forschung haben sich verschiedene Verfahren etabliert, wie Inhaltsvalidität gesichert werden kann, insbesondere eine strukturierte Befragung von Experten (oder informeller in Form ihrer Einbindung in verschiedenen Projektphasen), die Befragung von Testteilnehmern zu ihren kognitiven Prozessen (*cognitive diagnostic assessment;* Leighton & Gierl, 2007) im Zuge der Item-Bearbeitung (zum Beispiel in Form von *think-aloud interviews*; Ericsson & Simon, 1993) und die Orientierung der Itementwicklung an Dokumenten, die die zukünftig zu bewältigenden Anforderungen präzisieren (z.B. Bildungsstandards).

Mithilfe von *think-aloud interviews* wird untersucht, ob die Testteilnehmer tatsächlich jene kognitiven Fähigkeiten heranziehen, um die Items zu bearbeiten, die im zuvor entwickelten hypothetischen Modell angenommen und spezifiziert worden sind. Ein Beispiel für solche Validierungsstudien stellen Cui und Roberts (2013) vor. Verfahren zur Sicherung der inhaltlichen Passung von Testverfahren zu Dokumenten sind beispielsweise *alignment studies* zur Prüfung des Zusammenhangs zwischen schulischen Curricula, instruktionalem Lehrerhandeln und eingesetzten Test-Items (Webb, 2007; Davis-Becker & Buckendahl, 2013). Solche Verfahren sind insbesondere im Zuge von *high-stakes testing* wichtig geworden, bei denen weitreichende Konsequenzen aus Testleistungen gezogen werden sollen. Ein solches Vorgehen hat sich beispielsweise in den USA im Zuge der Umsetzung des „No Child Left Behind“-Act von 2001 etabliert.

Ein typisches Problem der Inhaltsvalidierung ist zum einen die kriteriale Festlegung von „Expertentum“ und zum anderen die Möglichkeit, dass unterschiedliche Experten zu unterschiedlichen Ergebnissen kommen. Insofern ist auf den Auswahlprozess und das Verfahren zur Aggregierung der Expertenmeinungen besondere Sorgfalt zu legen, soll die Expertenbefragung tatsächlich valide Ergebnisse produzieren. Neben die Schritte des Samplings von Probanden zur Durchführung einer Studie, von Aufgaben und Items für die Testentwicklung tritt im Zuge der Validierung also ein dritter Sampling-Prozess: der Experten und ihrer argumentativen Plausibilisierung. In jedem Falle gehört es zur Seriosität einer Studie, die Kriterien und Ergebnisse einer Expertenbefragung mit in die Berichterstattung aufzunehmen (Jonson & Plake, 1998).

Konstruktvalidität

So bedeutsam eine Sicherung von Kriteriums- und Inhaltsvalidität bis heute sind, so wenig zufriedenstellend ist ihre alleinige Anwendung bei der Messung hypothetischer Konstrukte. Diese Feststellung führte bereits in den 1950er Jahren zu einer Weiterentwicklung des Validitätskonzept im Auftrag der American Psychological Association (APA), indem der Begriff der „Konstruktvalidität“ entwickelt wurde. Als zentrale Arbeit kann in diesem Zusammenhang der Beitrag von Cronbach und Meehl (1955, p. 283) angesehen werden. Sie verweisen darauf, dass mit Tests erfasste Merkmale immer konstruierte Größen sind, die auf hypothetische Konstrukte bezogen sein sollten: „A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct.”

Damit wurde erstmalig eine explizite Verknüpfung von Theorie und Validität vorgenommen und die Leistungsdiagnostik in den hypothetisch-deduktiven Ansatz der Erkenntnisgewinnung eingeordnet (Frey, 2013). Das Besondere daran war, dass damit eine empirische Überprüfung von theoretisch abgeleiteten Hypothesen möglich wurde. Konstruktvalidität umfasst insofern 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, Übersetzung Johannes Hartig & Andreas Frey).

Drei empirische Strategien zur Stützung der Konstruktvalidität haben sich durchgesetzt (Hartig, 2013): die Prüfung der theoretisch angenommenen dimensionalen 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 im Sinne von Campbell und Fiske, 1959).

Die Grundidee einer Prüfung dimensionaler 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 (siehe auch im Folgenden Hartig, 2013). Beispiele hierfür sind, dass alle Items eines Tests ein gemeinsames Merkmal erfassen oder dass das Konstrukt eine mehrdimensionale Struktur aufweist, indem (Teil-)Kompetenzen voneinander abgrenzbar sind, also verschiedene Dimensionen darstellen. Solche Annahmen über die Ein- oder Mehrdimensionalität können in Modellen mit latenten Variablen, in denen die angenommenen Strukturen spezifiziert werden, geprüft werden, beispielsweise als Strukturgleichungsmodelle oder mehrdimensionale IRT-Modelle. Passt das Modell nicht zu den Testdaten, ist die Annahme widerlegt, dass der Test ein Konstrukt mit der angenommenen Struktur erfasst. Passt das Modell zu den Testdaten, unterstützt das die Annahmen über die Struktur – es ist jedoch noch offen, was der Test inhaltlich erfasst.

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 (Hartig, 2013). Die Testitems können dann auf einer „Konstruktlandkarte“ (*construct map*; Wilson, 2005) in der erwarteten Reihenfolge abgetragen werden, wobei die Schwierigkeit durch unterschiedliche Faktoren beeinflusst sein kann (Inhaltsbereich, kognitive Prozesse etc.; Embretson, 1983; Hartig & Frey, 2012).

Eine empirische Prüfung dieser Annahmen erfolgt, indem Hypothesen darüber formuliert werden, welche Charakteristika oder Inhalte von Aufgaben höhere Anforderungen an die zu messende Kompetenz darstellen, und dann die empirische Aufgabenschwierigkeit durch die angenommenen Faktoren zu erklären, z. B. in Form von Regressionsanalysen mit den Schwierigkeiten als abhängige Variable oder sogenannten 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 spezifizierten Aufgabenanforderungen erklären lassen. Gelingt dies nicht, sind die verwendeten Anforderungen für das Konstrukt irrelevant, was der angenommenen theoretischen Definition widerspricht.

Die Grundidee der Überprüfung, inwieweit sich das zu erfassende Konstrukt in ein nomologisches Netzwerk einfügen lässt, ist verwandt mit der oben beschriebenen Überprüfung externer bzw. krite-

rialer Validität (Hartig, 2013). Im Sinne von Cronbach und Meehl (1955) werden Annahmen darüber formuliert, mit welchen anderen Variablen das zu erfassende Konstrukt in welchem Zusammenhang stehen sollte. Das nomologische Netz besteht aus einem durch Korrespondenzregeln operationalisierten axiomatischen System und sogenannten empirischen Gesetzen (Frey, 2013). Es umfasst somit Elemente des Bereichs der Theorie und des Bereichs der Beobachtung und besteht in der Regel nicht aus einer einfachen Korrelationsmatrix, da alle zu prüfenden Zusammenhänge und Effekte theoriegeleitet begründet und einzeln zu prüfen sind.

Die empirische Prüfung erfolgt, indem die gerichteten oder ungerichteten Zusammenhänge des Testwertes mit anderen Variablen zum Beispiel manifest in Form von Korrelationsanalysen oder auf latenter Ebene in Form von Strukturgleichungsmodellen oder mehrdimensionalen IRT-Modellen untersucht werden (Hartig, 2013; siehe auch der Nachweis von konvergenter und diskriminanter Validität im Rahmen der Multitrait-Multimethod-Methode von Campbell und 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. Entspricht das Zusammenhangsmuster nicht dem theoretisch erwarteten, können die Testwerte nicht durch das angenommene Konstrukt erklärt werden oder die theoretischen Annahmen im nomologischen Netz sind (zumindest teilweise) falsch. Die Annahme der Validität einer Testwertinterpretation kann immer nur verworfen oder beibehalten, aber nicht abschließend belegt werden (Frey, 2013).

Verschiedene Strategien der Konstruktvalidierung schließen sich nicht gegenseitig aus, sondern können sich vielmehr ergänzen (Hartig, 2013). Welche Strategien sich zur Unterstützung der theoriebasierten Interpretation von spezifischen Testwerten empfehlen, hängt davon ab, wozu präzise theoretische Annahmen existieren. Es gilt also zu klären, ob 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 untersuchten Konstrukte häufig nicht fundiert genug, um Validitätsprüfungen durchführen zu können. Nach Cronbach und Meehl (1955, p. 284) ist in solchen Fällen folgende zentrale Voraussetzung nicht gegeben: „A construct has certain associated meanings carried in statements of this general character: Persons who possess this attribute will, in situation X, act in manner Y (with a stated probability). The logic of construct validation is invoked whether the construct is highly systematized or loose, used in ramified theory or a few simple propositions, used in absolute prepositions or probability statements. We seek to specify how one is to defend a proposed interpretation of a test: we are not recommending any one type of interpretation."

2 Empfehlungen der Fachverbände AERA, APA und NCME

Die Empfehlungen der Fachverbände zu Teststandards spiegeln die oben dargestellte Entwicklungsgeschichte wider (vgl. im Folgenden Frey, 2013). In der ersten Auflage der „Technical Recommendations for Psychological Tests and Diagnostic Techniques“ (APA, 1954) wird Validität noch als Eigenschaft eines Tests angesehen und es werden jene vier Kriterien von Gültigkeit unterschieden, die oben dargelegt sind: prädiktive Validität, konkurrente Validität, Inhaltsvalidität und Konstruktvalidität. Die Empfehlung der Fachverbände lautete, jene Validitätsart zu betrachten, die mit der intendierten Anwendung des Tests am besten korrespondierte. Mit der zweiten Auflage der nunmehr als „Standards for Educational and Psychological Tests and Manuals“ bezeichneten Empfehlungen (AERA, APA & NCME, 1966) werden prädiktive und konkurrente Validität wie im vorliegenden Beitrag zur Kriteriumsvalidität zusammengefasst. Die Empfehlung lautete nunmehr, mehr als einen Ansatz zu nutzen.

Die 1970er und 1980er Jahre sind dann von einem Wandel im Verständnis geprägt (Frey, 2013). Verstärkt sprechen sich die führenden Vertreter für eine gemeinsame Betrachtung der verschiedenen Validitätskriterien unter dem Dach der Konstruktvalidität aus (Cronbach, 1980; Guion, 1977; Messick, 1975, 1981). Allerdings wird zugleich davor gewarnt, dass Konstruktvalidität in der Praxis dann möglicherweise nur noch als „Mülleimerkategorie“ (Cronbach, 1980) verwendet werde, indem beliebige Korrelationen von Testergebnissen mit anderen Variablen ohne theoretische Einbettung als Beleg angesehen würden. Anastasi (1986) spricht in diesem Zusammenhang von „blind empirism“.

3 Verständnis von Validität heute

Die vierte Auflage der „Standards for Educational and Psychological Testing“ (APA, AERA & NCME, 1985) nahm den Wandel im Verständnis dessen, auf was sich eine Validierung richten sollte, umfassend auf. Validität wird seither definiert als „the approriateness, meaningfulness, and usefulness of specific inferences made from test scores“ (p. 9), womit die Gültigkeit der Testwertinterpretationen in den Vordergrund gerückt wurde. Inhalts-, Kriteriums- und Konstruktvalidität werden seither als verschiedene Arten der Evidenz für die Gültigkeit von Testwertinterpretationen beschrieben und nicht mehr als verschiedene Arten der Validität (Frey, 2013). Kane (1996, 2001) lenkte den Blick zudem weg von der reinen Absicherung durch formalisierte Theorien stärker hin zu einer überzeugenden, theoretisch fundierten argumentativen Stützung einer 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” (p. 13).

Die Diskussion blieb hierbei aber nicht stehen, sondern das Validitätsverständnis wurde von Kane im Anschluss an Arbeiten von Messick und Cronbach nochmals erweitert, indem neben eine messtheoretische Perspektive, in der Nachweise für die Präzision der Messung, ihre Konstruktvalidität und die Interpretation von Testscores gefordert wurden, darüber hinaus eine neue ethische Perspektive eingenommen wurde. In dieser werden die Fragen nach den Kosten und dem Nutzen von Testungen sowie nach ihren absichtlichen und unabsichtlichen Folgen bzw. Nebenwirkungen im Sinne einer *policy perspective* („should the test be used, does it have social value?“) aufgeworfen. Messick (1989) hatte diese Form der Validierung als *consequential validity* bezeichnet. Diese Perspektive ist sicher

die umstrittenste – nicht nur, weil sie schwierig empirisch zu prüfen ist und sich nicht gut in das Rational der Konstruktvalidität einpasst. Sie nimmt die Testentwicklung auch ein Stück weit in Haftung für den späteren Einsatz ihrer Tests. Was Kane (2013) damit zumindest erreichen will, ist eine Unterstützung des Testeinsatzes in der Praxis durch die ursprünglichen Testentwickler – ggf. auch noch lange nachdem diese das Projekt beendet haben. Er führt dafür technische Argumente an – die Testentwickler haben einerseits die größte Expertise in diesem Bereich –, aber andererseits auch ganz fundamentale soziale: „we all are expected to take responsibility for what we do“ (p. 62)

Dass Kane und andere (zum Beispiel Heubert & Hauser, 1999, für das Committee on Appropriate Test Use des National Research Council) auf dem Validitätskriterium der mit Tests verbundenen Konsequenzen und Entscheidungsprozesse beharren, hat möglicherweise mit den zahlreichen negativen Erfahrungen zu tun, die in den USA mit Testprogrammen gemacht wurden. Kane verweist in seinem aktuellen umfangreichen Beitrag zur Validitätsdebatte jedenfalls mehrfach auf den No-Child-Left-Behind-Act sowie darauf, 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” (Kane, 2013, p. 52). Er schlussfolgert: “The accountability program is an educational intervention, and a serious evaluation of an accountability program would require an evaluation of both intended and unintended outcomes.” (p. 54)

Allerdings hat sich gegen diese sehr weite Perspektive mittlerweile auch Widerspruch formiert. Wissenschaftler wie Borsboom, Mellenbergh und van Heerden (2004) oder Scriven (2010) betonen, dass ein engeres Konzept der Validierung notwendig sei, um den Begriff nicht zu überladen und ihm damit seine Eindeutigkeit zu nehmen. Präzision wird hier als ein Qualitätsmerkmal angesehen. Auch die Konsequenzen eines Testeinsatzes einschließen zu müssen, mache Validierungen praktisch undurchführbar. Zudem könne eine missbräuchliche Verwendung niemals ausgeschlossen werden.

Borsboom et al. (2004) führen neben pragmatischen auch wissenschaftstheoretische Argumente an, warum sie eine Validierung von Instrumenten und nicht nur von Testwertinterpretationen für bedeutsam halten. Ihr Ausgangspunkt ist, dass die von einem Test erfasste Eigenschaft unabhängig vom Testverfahren existiere und ein Test dann valide sei, “if variation in the attribute causes variation in the test scores” (p. 1067). Sie proklamieren also zum einen eine beobachtungsunabhängige Existenz von Eigenschaften – an einigen Stellen sprechen sie auch von truth – und zum anderen eine explizit kausale Beziehung zwischen Konstrukt und Testwert (siehe auch Borsboom, Cramer, Kievit, Zand Scholten & Franic, 2009). Borsboom et al. sperren sich dabei nicht grundsätzlich gegen die Berücksichtigung der von Kane diskutierten Kriterien als Gütekriterien empirischer Forschung, aber nur im Sinne einer generellen Qualität dieser, nicht als Aspekt des Validitätsbegriffs.

Dass es ihnen tatsächlich um das Verhältnis von Validität und „Wahrheit“ (*truth*) geht, bekräftigen Boorsboom und Markus (2013) in ihrer Replik auf Kane (2013), in der sie die im 17. und 18. Jahrhundert vertretene, später dann aber widerlegte Phlogiston-Theorie als Beispiel nehmen. Verschiedene Chemiker dieser Zeit nahmen die Existenz eines flüchtigen Elements „Phlogiston“ an, das Verbrennungsprozesse befördere. Borsboom und Markus (2013) kritisieren nun, dass nach dem Validitätskonzept von Kane (2013) diese Interpretation vorliegender Testwerte als validiert gegolten hätte, obwohl sie objektiv fehlerhaft war. Kane (2013) reagiert darauf in seiner Replik wiederum wie folgt: „I would say that within the context of 18th century science, the interpretation of certain measures as indicators of the amount of phlogiston in a body was reasonable in light of prevailing theory and

all available evidence and therefore was valid in that context. In the context of current theory, interpretations in terms of phlogiston do not make sense and would not be considered valid. The phlogiston theory has been overturned, and interpretations based on the discredited theory have been discarded. Scientific theories and interpretations based on these theories are fallible and subject to revision." Kane wendet sich auch gegen den Wahrheitsanspruch von Konstrukten, den er in der pädagogisch-psychologischen Forschung für unangemessen hält. In der Tat ist zu fragen, ob in komplexen Gebieten wie der Kompetenzmessung Aussagen getroffen werden können, die im axiomatischen Sinne als „wahr" angesehen werden können.

4 Schlussfolgerungen für das Validitätsverständnis im Forschungsprogramm KoKoHs

Zusammenfassend kann festgehalten werden, dass es im Zusammenhang der Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor dem aktuellen Stand der Forschung entsprechen würde, wenn die Validität verschiedener *Interpretationen* von Testergebnissen betrachtet würde, statt von „der Validität eines Tests" zu sprechen. Im Zuge der Validierung gilt es für die 23 Verbundprojekte daher im ersten Schritt zu spezifizieren, auf welche Interpretation eines Testergebnisses sich die intendierte Validierung bezieht (Hartig, 2013). Verschiedene Interpretationen können sich zum Beispiel beziehen 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) Erklären eines Testwertes und auf das Treffen weiterführender Entscheidungen als Konsequenz aus dem Testergebnis (Kane, 2001).

Entsprechend ist die Validierung eines in KoKoHs entwickelten Tests „kein immer gleiches Routineverfahren, sondern erfolgt durch theoriegeleitete Forschung, mit der unterschiedliche Interpretationen eines Testergebnisses legitimiert oder auch falsifiziert werden können" (Hartig, Frey & Jude, 2012, S. 145). Ob tatsächlich auch langfristige intendierte und nicht-intendierte Folgen und Nebenwirkung eingeschlossen werden sollten, muss an dieser Stelle offen bleiben.

Literatur

Albino, J. E., Young, S. K., Neumann, L. M., Kramer, G. A., Andrieu, S. C., Henson, L., Horn, B. & Hendricson, W. D. (2008). Assessing Dental Students' Competence: Best Practice Recommendations in the Performance Assessment Literature and Investigation of Current Practices in Predoctoral Dental Education. *Journal of Dental Education, 72,* 1405-1435.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–15.

Blömeke, S., Busse, A., Suhl, U., Kaiser, G., Benthien, J., Döhrmann, M. & König, J. (eingereicht). Entwicklung von Lehrpersonen in den ersten Berufsjahren: Längsschnittliche Vorhersage von Unterrichtswahrnehmung und Lehrerreaktionen durch Ausbildungsergebnisse. *Zeitschrift für Erziehungswissenschaft*.

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

Borsboom, D., Cramer, A., Kievit, R., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. Lissitz (Ed.), *The concept of validity* (pp. 135–170). Charlotte, NC: Information Age Publishers.

Borsboom, D. & Markus, K. A. (2013). Truth and Evidence in Validity Theory. *Journal of Educational Measurement*, 50, 110–114.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement: Measuring Achievement Over a Decade,* 5, 99–108.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

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

Cui, Y. & Roberts, M. R. (2013). Validating Student Score Inferences With Person-Fit Statistic and Verbal Reports: A Person-Fit Study for Cognitive Diagnostic Assessment. *Educational Measurement: Issues and Practice*, 32, 34–42.

Davis-Becker, S. L. & Buckendahl, Ch. W. (2013). A Proposed Framework for Evaluating Alignment Studies. *Educational Measurement: Issues and Practice*, 32, 23–33.

Deutsches PISA-Konsortium (Hrsg.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske und Budrich.

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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

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

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.): *Test- und Fragebogenkonstruktion*. 2. Auflage. (S. 143-171). Berlin: 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.

Heubert, J. P., & Hauser, M. H. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academies Press.

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

Kane, M. (2002b). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31–41.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement Spring 2013, Vol. 50, No. 1, pp. 1–73

Kane, M. T. (2013). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement*, 50, 115–122.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007a). *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge, UK: Cambridge University Press.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, 10(9), 9–20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

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.

Miller, G. E. (1990). The Assessment of Clinical Skills/Competence/Performance. Academic Medicine. *Journal of the Association of American Medical Colleges*,65, 63–67.

Nerdinger, F., Blickle, G. & Schaper, N. (2011). *Lehrbuch Arbeits- und Organisationspsychologie* (2. Aufl.). Heidelberg, Berlin, New York: Springer.

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.

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

Scriven, M. (2010). Rethinking evaluation methodology. *Journal of MultiDisciplinary Evaluation*, *6*(13), i-ii. 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.

Vogelsang, C. & Reinhold, P. (2012). Wissen und Handeln angehender Physiklehrkräfte. In Höttecke, D. (Hrsg.), *Konzepte fachdidaktischer Strukturierung für den Unterricht – Gesellschaft für Didaktik der Chemie und Physik – Jahrestagung in Oldenburg 2011*. Münster: LIT.

Wass, V., Van der Vluten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet*, 357, 945–949.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7–26.

Wilson, M. (2005). *Constructing measures. An item response modelling approach*. Mawah: Lawrence Erlbaum Associates.

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.

Bisher erschienen:

KoKoHs Working Papers, 1

Blömeke, S. & Zlatkin-Troitschanskaia, O. (2013). Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs [Modeling and Measuring Competencies in Higher Education: Aims, theoretical framework, design, and challenges of the BMBF-funded research program KoKoHs] (KoKoHs Working Papers, 1). Berlin & Mainz: Humboldt-Universität & Johannes Gutenberg-Universität.

KoKoHs Working Papers, 2

Blömeke, S. (2013). Validierung als Aufgabe im Forschungsprogramm “Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor” [The task of validation in the research program „Modeling and Measuring Competencies in Higher Education”] (KoKoHs Working Papers, 2). Berlin & Mainz: Humboldt-Universität & Johannes Gutenberg-Universität.