Schaper, Nicolas

Validitätsaspekte von Kompetenzmodellen und -tests für hochschulische Kompetenzdomänen

Musekamp, Frank [Hrsg.]; Spöttl, Georg [Hrsg.]: Kompetenz im Studium und in der Arbeitswelt. Nationale und internationale Ansätze zur Erfassung von Ingenieurkompetenzen. Frankfurt, M. : Lang 2014, S. 21-48. - (Berufliche Bildung in Forschung, Schule und Arbeitswelt; 12)

Frank Musekamp / Georg Spöttl (Hrsg./eds.)

Kompetenz im Studium und in der Arbeitswelt

Competence in Higher Education and the Working Environment

Nationale und internationale Ansätze zur Erfassung von Ingenieurkompetenzen

National and International Approaches for Assessing Engineering Competence

Berufliche Bildung in Forschung, Schule und Arbeitswelt

Vocational Education and Training: Research and Practice

Herausgegeben von Matthias Becker und Georg Spöttl

Band 12

PETER LANG EDITION

Gute Lehre an Hochschulen hat in den vergangenen Jahren erheblich an Bedeutung gewonnen. Dies gilt angesichts hoher Studienabbruchquoten insbesondere in den Ingenieurswissenschaften. Um die Effekte guter Lehre auf die Lernergebnisse bei den Studierenden zu erfassen, fehlen jedoch bislang empirisch abgesicherte Instrumente. Dieser Band stellt aktuelle konzeptionelle und empirische Arbeiten vor und beleuchtet sie aus methodischer Sicht sowie mit Blick auf die didaktische Verwertung in der Hochschullehre.

Good teaching at universities has considerably gained importance within the last years. This is especially relevant with regard to the high drop-out rates, above all in engineering sciences. At the moment, however, there is a lack of empirically valid instruments for the assessment of the impact of good teaching on the students' learning results. This volume presents current conceptual and empirical works with a focus on methodology and their didactical application in university teaching.

Frank Musekamp ist Wissenschaftlicher Mitarbeiter am Institut Technik und Bildung (ITB) der Universität Bremen.
Georg Spöttl, Dr. Dr. h. c. ist Professor für Didaktik und Leiter der Abteilung Arbeitsprozesse und berufliche Bildung am Institut Technik und Bildung (ITB) der Universität Bremen.

Frank Musekamp is research associate at the Institute Technology and Education (IT-B) of the University of Bremen.
Georg Spöttl, Dr. Dr. h. c., is professor for didactics and director of the Department Work Processes and Vocational Education at the Institute Technology and Education (IT-B) of the University of Bremen.

www.peterlang.com

Kompetenz im Studium und in der Arbeitswelt
Competence in Higher Education and the Working Environment

Berufliche Bildung in Forschung, Schule und Arbeitswelt

Vocational Education and Training: Research and Practice

Herausgegeben von Matthias Becker und Georg Spöttl

Band 12

Zur Qualitätssicherung und Peer Review der vorliegenden Publikation

Die Qualität der in dieser Reihe erscheinenden Arbeiten wird vor der Publikation durch externe, von der Herausgeberschaft benannte Gutachter im Blind Verfahren geprüft. Dabei ist der Autor der Arbeit den Gutachtern während der Prüfung namentlich nicht bekannt.

Notes on the quality assurance and peer review of this publication

Prior to publication, the quality of the work published in this series is blind reviewed by external referees appointed by the editorship. The referees are not aware of the author's name when performing their review.

Frank Musekamp / Georg Spöttl (Hrsg./eds.)

Kompetenz im Studium und in der Arbeitswelt

Competence in Higher Education and the Working Environment

Nationale und internationale Ansätze
zur Erfassung von Ingenieurkompetenzen

National and International Approaches
for Assessing Engineering Competence

PETER LANG
EDITION

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Library of Congress Cataloging-in-Publication Data
Kompetenz im Studium und in der Arbeitswelt : nationale und internationale Ansätze zur Erfassung von Ingenieurkompetenzen = Competence in higher education and the working environment : national and international approaches for assessing engineering competence / Frank Musekamp, Georg Spöttl, Hrsg.-eds.
pages cm. – (Vocational education and training : research and practice, 1865-844x ; Band 12)
Parallel title: Competence in higher education and the working environment
Parallel text in German and English.
ISBN 978-3-631-65104-9
1. Engineering–Study and teaching. 2. Engineering–Vocational guidance.
3. Engineers. I. Musekamp, Frank, 1978- editor. II. Spöttl, Georg, editor.
III. Title: Competence in higher education and the working environment.
TA157.K646 2014
620.0071–dc23
2015001869

Dieses Buch wird aus Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01PK11012A gefördert.
Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

ISSN 1865-844X
ISBN 978-3-631-65104-9 (Print)
E-ISBN 978-3-653-04168-2 (E-Book)
DOI 10.3726/978-3-653-04168-2

Diese Publikation wurde begutachtet.

www.peterlang.com

Niclas Schaper

Validitätsaspekte von Kompetenzmodellen und -tests für hochschulische Kompetenzdomänen

Vor dem Hintergrund von Problemen bei der Validitätsbestimmung im Kontext von Ansätzen zur Kompetenzmodellierung und -messung im Hochschulsektor wird begründet, warum ein erweitertes Validitätsverständnis, dass sich an dem Konzept der argumentbasierten Validierung nach Messick bzw. Kane orientiert, bei entsprechenden Modellierungs- und Messansätzen für sinnvoll und angemessen gehalten wird. Die verschiedenen Validierungsaspekte des Messick'schen Ansatzes: (1) Inhaltliche Validität, (2) Kognitive Validität, (3) Strukturelle Validität, (4) Verallgemeinerbarkeit, (5) Externe Validität und (6) Konsequentielle Validität werden anschließend vorgestellt und hinsichtlich ihrer Relevanz für die Validierung von Kompetenzmodellen und -tests erörtert sowie anhand von Beispielen verdeutlicht.

Based on a description of problems concerning the analysis of validity of approaches of competence modelling and measurement it is substantiated why it is reasonable and adequate to refer to the concept of "argument based validity" according to Messick resp. Kane in this context. Following, the different validity aspects of the Messick approach are introduced: (1) content validity, (2) cognitive validity, (3) structural validity, (4) generalizability, (5) external validity, and (6) consequential validity. These validity aspects are discussed concerning their relevance for this context and examples of an argument based analysis of validity of competence models and tests are given.

1 Probleme der Validitätsbestimmung bei der Kompetenzmodellierung und -messung im Hochschulsektor

Den Kompetenzerwerb bzw. die Kompetenzentwicklung von Akademikern im Kontext verschiedener hochschulischer Bildungsinstitutionen zu erfassen, stellt eine theoretische und methodische Herausforderung dar. Zum einen liegen kaum theoretische Vorarbeiten bezüglich des Kompetenzerwerbs in akademischen Kontexten vor (Schaper, 2012). Zum anderen ist eine valide und zuverlässige Modellierung und Erfassung akademisch vermittelter Kompetenzen sowie ihrer Bedingungen und Wirkungen aufgrund ihrer Multidimensionalität und -kausalität mit hohen Ansprüchen an die Forschungsmethodik verbunden (Blömeke & Zlatkin-Troitschanskaia, 2011). Das vorhandene Forschungsdefizit ist zudem in Teilen auf die besondere Komplexität zurückzuführen, die akademisch erworbene Kompetenzen von Studierenden und Promovierenden aufgrund der Vielfalt an Studienmodellen,

Ausbildungsstrukturen und Lehrangeboten auszeichnet. Die Wahl angemessener Kriterien, anhand derer der Kompetenzerwerb eingeschätzt werden kann (z. B. beruflicher Erfolg oder Bewältigung ausgewählter Berufsanforderungen), stellt ebenfalls ein bisher nur unzureichend gelöstes Problem dar. Berufliche Einsatzfelder und Anforderungen an Akademiker sind in vielen Fachdomänen (z. B. geistes- oder sozialwissenschaftlichen Fächern) nur schwer erfassbar bzw. definierbar und unterliegen einer dynamischen Wandlung. Aber auch Fragen wie akademische Kompetenzen bezüglich ihrer inhaltlichen Validität und Konstruktvalidität überprüft werden können, sind allenfalls in Ansätzen gelöst (Blömeke, 2013).

Eine zentrale Frage in diesem Zusammenhang betrifft außerdem das Validitätsverständnis, das bei der Modellierung und Messung akademischer Kompetenzen zugrunde gelegt wird. Das klassische Validitätsverständnis, das darauf beruht, dass Validität als statistisch quantifizierbare Eigenschaft eines Tests angesehen wird, greift gerade bei der Beurteilung und Überprüfung der Aussagefähigkeit eines Kompetenztests zu kurz (Blömeke, 2013). Bei dem klassischen Validitätsverständnis stehen die statistischen Zusammenhänge mit konstruktnahen oder kriteriumsbezogenen Variablen im Vordergrund (vgl. z. B. Lienert & Raatz, 1994). Solche Kennwerte sagen allerdings wenig aus über die Angemessenheit der theoretischen Annahmen, die dem Test zugrunde liegen, die Konvergenz von Testleistungen und dem theoretischen Kompetenzmodell oder auch die Passung von Kompetenzmodell und gewähltem psychometrischem Modell. Weiterhin greifen die klassischen Validitätsbetrachtungen zu kurz, wenn es um die Verallgemeinerbarkeit der diagnostischen Ergebnisse über bestimmte Aufgaben- und Personengruppen hinaus geht oder um die Angemessenheit der praktischen Schlussfolgerungen, die aus den diagnostischen Ergebnissen einer Person gezogen werden. Die genannten Validitätsaspekte sind insbesondere für die Modellierung und Messung von Kompetenzen in akademischen Ausbildungskontexten von Bedeutung. Bezüglich der Entwicklung valider Modelle und Messinstrumente zur Diagnose akademischer Kompetenzen besteht generell noch erheblicher Forschungsbedarf. Dieser wird zurzeit in einem umfangreichen Forschungsprogramm des BMBF in 23 Verbundprojekten unterschiedlicher Fachrichtungen angegangen, um Grundlagen für eine theoretisch fundierte und valide Messung akademischer Kompetenzen von Studierenden zu erarbeiten (Blömeke & Zlatkin-Troitschankaja, 2013). Damit soll eine grundlagenorientierte Kompetenzforschung im Bereich der Hochschulen in Deutschland voran gebracht werden und die Anschlussfähigkeit an internationale Forschungsansätze der empirischen Bildungsforschung gewährleistet werden.

Im Folgenden wird zunächst kurz begründet, warum ein erweitertes Validitätsverständnis gemäß dem sog. argument based approach zur Validierung

von Kompetenzmodellen und -tests für sinnvoll gehalten wird. In weiteren Kapiteln werden vor diesem Hintergrund die verschiedenen Validierungsaspekte des Messick'schen Ansatzes vorgestellt und hinsichtlich ihrer Relevanz für die Validierung von Kompetenzmodellen und -tests erörtert sowie anhand von Beispielen verdeutlicht.

2 „Argument based approach" als angemessene Validierungsstrategie für Kompetenzmodelle und -tests

Validität ist in der psychologisch-pädagogischen Diagnostik – neben der Objektivität und der Reliabilität – ein zentrales Gütekriterium eines Testverfahrens zur Erfassung eines psychologischen Merkmals. Es wird oft mit der Formulierung beschrieben, dass Validität ein Kriterium dafür darstellt, „ob ein Test tatsächlich das misst, was er messen soll" (Bortz & Döring, 2006). Validität bezieht sich also darauf, ob ein Test eine gewisse ‚Gültigkeit' besitzt, um Aussagen über die Ausprägung eines bestimmten Merkmals einer Person treffen zu können. Sie bezieht sich daher nicht auf ein Testverfahren ‚an sich', sondern auf Aussagen und Interpretationen, die auf der Basis von Ergebnissen aus diesem Verfahren vorgenommen werden. Dieses Verständnis von Validität bezieht sich also auch auf die mit einem Testverfahren verbundenen theoretischen Annahmen, die Interpretationen von Ergebnissen und die Schlussfolgerungen, die aus diesen Ergebnissen gezogen werden (Messick, 1995). Um die Validität eines Verfahrens einschätzen zu können, müssen daher auch Evidenzen bzw. Erkenntnisse ermittelt werden, die die Interpretation von Testergebnissen und die daraus gezogenen Schlussfolgerungen plausibel stützen.

Es müssen also Argumente dafür gefunden werden, ob eine Interpretation als plausibel, sinnvoll bzw. angemessen angenommen werden kann. Diese Argumente sollten sich sowohl auf empirische Evidenzen als auch auf theoretisch-rationale Begründungen und Prinzipien beziehen. Dabei können verschiedene Testverfahren bezüglich unterschiedlicher Aussagen unterschiedliche Grade der Validität besitzen, also unterschiedlich ‚valide sein'. Kane (2013, S. 3) drückt diesen Zusammenhang folgendermaßen aus: „Interpretations and uses that make sense and are supported by appropriate evidence are considered to have high validity (or for short, to be valid), and interpretations or uses that are not adequately supported, or worse, are contradicted by the available evidence are taken to have low validity (or for short, to be invalid). The scores generated by a given test can be given different interpretations, and some of these interpretations may be more plausible than others".

Dieses Verständnis von Validität als Argumente für die ‚Angemessenheit einer Testinterpretation' wird als „argument based approach to validation",

der Prozess des Findens und der Prüfung solcher Argumente dementsprechend als Validierung bezeichnet (Kane, 1992 bzw. Messick, 1995). Demnach ist es auch sinnvoll, von der Validität eines Kompetenzmodells zu sprechen, wie es in der deutschsprachigen Lehrerbildungsforschung häufig getan wird (z. B. Borowski et al., 2010; Schaper, 2009), da sich Validitätsargumente im Sinne von Interpretationen auch auf theoretische Annahmen und Begründungen beziehen.

Neben diesem übergeordneten Validitätsverständnis existieren verschiedene Konzeptionen, Klassen und Begriffe von Validität, die sich im Wesentlichen darin unterscheiden, welche Art von Evidenzen sie zu einem Validitätsargument beitragen, welche davon überhaupt als ‚angemessen' angenommen werden und mit Hilfe welchen Vorgehens solche Evidenzen ermittelt werden können. Diese Konzeptionen basieren dabei teilweise auf unterschiedlichen, historisch gewachsenen Forschungstraditionen und beinhalten daher auch unterschiedliche implizite Annahmen. Bspw. wird in der pädagogisch-psychologischen Forschung häufig auf das Konzept der Konstruktvalidität zurückgegriffen (Cronbach & Meehl, 1955). Dabei wird versucht, Argumente für die Validität eines Verfahrens dadurch zu finden, dass „aus den Annahmen für das [zu messende] Konstrukt Vorhersagen abgeleitet werden, wie Messwerte des Konstrukts mit anderen Variablen zusammenhängen sollten" (Hartig & Jude, 2007). Diese Vorhersagen bilden ein zusammenhängendes nomologisches Netzwerk von Aussagen, die empirisch dadurch überprüft werden können, indem neben Messungen des eigentlich interessierenden Konstrukts mit dem zu prüfenden Verfahren zeitgleich Messungen mit anderen Verfahren vorgenommen werden, die Variablen erfassen, mit denen das Konstrukt gemäß des Netzwerks zusammenhängt (a.a.O.).

Grundsätzlich gilt, dass auch bei Kompetenzmodellen und -tests die klassischen Gütekriterien ihre Relevanz behalten (vgl. z. B. Schaper, 2009); dies gilt sowohl für inhalts-, konstrukt- und kriterienbezogene Validitätsaspekte. Der Argument basierte Ansatz richtet den Fokus bei der Validierung allerdings auf zusätzliche Aspekte, die für die Validität von Kompetenzmodellen und -tests eine zentrale Bedeutung haben, jedoch häufig in diesem Zusammenhang vernachlässigt werden. Dies betrifft z. B. die Frage, welche theoretischen Annahmen dem Kompetenztests in Bezug auf den Zusammenhang von bestimmten Kompetenzfacetten und ihre Relevanz für die Handlungsbefähigung für bestimmte Professionskontexte bzw. -aufgaben zugrunde liegen und ob diese vor dem Hintergrund kognitionspsychologischer oder handlungstheoretischer Konzepte tatsächlich plausibel und angemessen sind (z. B. die Frage, inwieweit Wissen tatsächlich eine relevante Voraussetzung für effektives Handeln darstellt und wie man sich den Zusammenhang von Wissen und Handeln vorstellen kann; siehe hierzu z. B. Vogelsang, 2014 für die Validierung eines Kompetenztests in der Physiklehrerausbildung).

Auch die Frage, ob die Testitems des Kompetenztests tatsächlich auch die Leistung bzw. die kompetenzrelevanten Leistungsvoraussetzungen erfassen, die gemäß einem (theoretisch fundierten) Kompetenzmodell relevant sind, ist insbesondere bei komplexen szenario- bzw. simulationsgestützten Erfassungsformaten nicht ohne entsprechende kognitive Aufgabenanalysen sicher beantwortbar und daher unter Validitätsgesichtspunkten gesondert zu prüfen (siehe hierzu auch Kap. 3.2). Weiterhin ist bei Kompetenztests oftmals die Frage, für welches Aufgaben- bzw. Fähigkeitsspektrum der Test Aussagen erlaubt, da die Items bzw. Leistungsanforderungen situationsspezifisch ausgestaltet sind und damit sowohl theoretisch als auch empirisch geklärt werden sollte, auf welche Anforderungen bzw. Aufgaben die gezeigten Leistungen verallgemeinert werden können oder auch auf welche Leistungen in anderen Bereichen extrapoliert werden kann. Schließlich werden auf der Basis der Testergebnisse bei Kompetenztest auch Entscheidungen über das Erreichen bestimmter Bildungsziele bzw. die Zertifizierung von erfolgreich bestandenen Bildungsmaßnahmen sowie den Bedarf von Fördermaßnahmen getroffen. Es gilt somit auch theoretische und empirische Evidenzen über die Angemessenheit solcher Schlüsse zu generieren, die bedeutungsvollen individuellen und sozialen Konsequenzen der Testanwendung verbunden sind. Die angeschnittenen Fragen bzw. Probleme bei der Entwicklung und Verwendung von Kompetenztests betreffen Validitätsaspekte, die im Rahmen der klassischen Validitätsanalysen allenfalls am Rande betrachtet werden. Der Ansatz von Messick (1995) bzw. Kane (2001) greift diese Lücken der Validitätsbetrachtung allerdings gezielt auf und vermittelt Hinweise, wie sie effektiver bei der Testkonstruktion und -überprüfung berücksichtigt werden können.

3 Aspekte eines umfassenden Validitätsverständnisses nach Messick und Kane

Die Grundidee von Messick (1995), die Validität einer diagnostischen Messung nicht allein als einen numerischen Koeffizienten zu betrachten, sondern vielmehr als theoretisch und empirisch fundiertes Argument für die Gültigkeit von Testwertinterpretationen, folgen mittlerweile eine ganze Reihe von Vereinigungen der pädagogisch-psychologischen Forschung (z. B. American Educational Research Association (AERA) oder American Psychological Association (APA)).

Validität ist für Messick (1995) bzw. Kane (2001) – wie oben bereits ausgeführt wurde – die allgemeine Bewertung des Grades, in dem empirische und theoretische Evidenz für Angemessenheit von Interpretationen und Handlungen, die auf diagnostischen Messungen beruhen, vorliegt. Das heißt, dass vor allem die Bedeutung, die Interpretation und die daraus abgeleiteten

Handlungen von diagnostischen Ergebnissen valide sein müssen. Das Ausmaß, in dem die Bedeutung von Testwerten und den daraus abgeleiteten Schlussfolgerungen über verschiedene Personen, Gruppen und Situationen stabil ist, ist jedoch eine nicht abschließend beantwortbare empirische Frage.

In seinem Modell greift Messick (1989) die klassische Einteilung von Validität in Inhalts-, Konstrukt- und Kriteriumsvalidität auf und entwickelt sie zu einem umfassenderen Validitätsverständnis weiter, das anhand von sechs Aspekten definiert wird. Nach Messick (1989) stellen diese sechs Aspekte generelle Validitätskriterien bzw. Standards für alle diagnostischen Messungen im Bereich von Bildung dar. Validität ist daher zu verstehen als Argument für die Gültigkeit von Testwertinterpretationen auf Grundlage von Evidenzen bzw. Erkenntnissen in diesen sechs Bereichen. Der Ansatz von Messick (1989, 1995) markiert damit in besonderer Weise die Abwendung vom klassischen Validitätsverständnis als einer Reihe von Eigenschaften eines Messinstruments und eine Hinwendung zu einem integrativen Konzept von Konstruktvalidität als fortwährendem Prozess der argumentativen und empirischen Verteidigung miteinander verbundener Validitätsaspekte. Vor dem Hintergrund dieses veränderten Validitätskonzepts wird im Folgenden Validität als die Gesamtbewertung der theoretischen Argumente und empirischen Evidenzen für die Angemessenheit einer Leistungsmessung und zwar sowohl ihrer Interpretation als auch der Konsequenzen ihrer Anwendung verstanden. Diese Sicht von Validität erstreckt sich sowohl auf die Grundlagenforschung zu diagnostischen Verfahren als auch auf deren Anwendung in der Praxis und ist daher angemessen für die Analyse entsprechender Ansätze der Kompetenzmodellierung und -messung, die dieses Einsatzspektrum für sich beanspruchen (Klieme & Leutner 2006). Nach Messick (1995) werden folgende sechs Validitätsaspekte unterschieden und beschrieben:

1) Inhaltliche Validität: Curriculare und theoretische Absicherung des modellierten Bereichs (content aspect)
2) Kognitive Validität: Passung der kognitiven Prozesse bei der Kompetenzerfassung zum postulierten theoretischen Kompetenzmodell (substantive aspect)
3) Strukturelle Validität: Passung von theoretischem Kompetenzmodell und gewähltem psychometrischem Messmodell (structural aspect)
4) Verallgemeinerbarkeit: Angemessenheit einer über die Aufgaben- und Personengruppe hinausgehenden Interpretation (generalizability aspect)
5) Externe Validität: Angemessenheit mit Blick auf konvergente, diskriminante und prädiktive Zusammenhänge mit anderen Konstrukten (external aspect)

Die genannten Validitätsaspekte erlauben eine systematische, sämtliche Schritte einer Kompetenzmodellierung durchdringenden Analyse von

Validität, die dabei hilft, die für jeden Schritt spezifischen Bedrohungen der Validität zu erkennen und zu bewältigen. Die genannten Validitätsaspekte sind dabei nicht unabhängig, sondern bedingen sich gegenseitig (z. B. die kognitive Validität und der Aspekt der Verallgemeinerbarkeit der Testergebnisse; wenn die Items nicht sorgfältig in Anlehnung an bestimmte theoretische Annahmen bzw. in Abhängigkeit von der Konstruktbedeutung operationalisiert werden, können auch keine weitreichenden Schlüsse über die Verallgemeinerbarkeit der erfassten Kompetenzen gezogen werden). Validität findet somit ihren Ausdruck nicht im additiven Vorliegen einzelner Eigenschaften, sondern in der Passung der Validitätsaspekte auf verschiedenen Ebenen, die im Prozess der Konstruktion und Anwendung eines Verfahrens von Bedeutung sind und miteinander interagieren.

Validität wird in der Regel als zentrales Qualitätskriterium herausgestellt. Sie ist aber nicht reduzierbar auf in Kennwerten ausdrückbare Eigenschaften eines Messinstrumentes, sondern stellt einen fortwährenden argumentativen Prozess dar. Dieser umfasst insbesondere auch die (tatsächlichen oder potenziellen) individuellen und sozialen Konsequenzen beim Einsatz eines (Kompetenz-) Tests: „[…] to appraise how well a test does its job, one must inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but also at the same time consistent with other social values“ (Messick, 1995, S. 165). Dies hat, angesichts ihrer Bedeutung in der Steuerung von Bildungssystemen, für moderne Kompetenzmodellierungen besonderes Gewicht. Im Folgenden werden die beschriebenen sechs Validitätsaspekte auf ausgewählte Ansätze der Modellierung von Kompetenzen bezogen. Die folgenden Ausführungen zu den verschiedenen Validitätsaspekten nach Messick (1995) lehnen sich eng an die Darstellung des Ansatzes von Leuders (2014) an.

3.1 Inhaltliche Validierung

Inhaltsvalidität liegt bei Evidenzen über die Relevanz eines gemessenen (Test-)Inhaltes und dessen Repräsentativität für das interessierende Konstrukt im Allgemeinen vor (Messick, 1989). Letzteres bezieht sich auf die Frage, ob die Test- und Aufgabeninhalte den interessierenden Merkmals- oder Verhaltensbereich, der das zu messende Konstrukt definiert, gut repräsentieren. Inhaltsvalidität wird in der Regel mithilfe so genannter Expertenbefragungen erfasst, wobei die Experten gebeten werden, die Inhalte eines Tests bezüglich Relevanz und Repräsentativität zu bewerten (vgl. für ein Beispiel Jenßen et al., 2014). Dies kann aber auch durch eine „Delphi-Befragung“, bei der alle Experten miteinander diskutieren, ob ein Item geeignet ist, realisiert werden (Kunina-Habenicht et al., 2012). Ein typisches Problem in diesem Zusammenhang ist zum einen die Festlegung von „Expertentum“ anhand von entsprechenden Kriterien 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. Jenßen et al. (2014) haben bei der Durchführung und Auswertung der Expertenbefragung zur Inhaltlichen Validierung von Testitems fünf Schwierigkeiten bzw. Probleme identifiziert: (a) Beurteilungsfehler durch die gleiche (falsche) Vorstellung des Konstrukts von Testkonstrukteuren und Experten; (b) einseitige Fehlvorstellung des Konstrukts auf Seiten der Testkonstrukteure oder Experten; (c) Transparenz von Projekthintergrund und -vorgehen; (d) Präzision der Kodieranweisungen und (e) Einbindung von Experten in weitere Phasen der Testkonstruktion. Im Beitrag werden zu jedem Problem möglich Ursachen bzw. Hintergründe aber auch Lösungsansätze zur Problembewältigung diskutiert.

Bei der Beurteilung der Inhaltsvalidität eines Tests sind u. a. folgende Quellen mangelnder Validität zu berücksichtigen: (a) Unterrepräsentation des zu messenden Konstrukts und (b) konstruktirrelevante Testvarianz (Messick, 1989). Im Fall einer Unterrepräsentation des Konstrukts ist ein Test zu eng gefasst und lässt wichtige Facetten und/oder Dimensionen des Konstrukts unberücksichtigt. Ein Test zur Messung sprachlicher Kompetenzen in Orientierung am Europäischen Referenzrahmen wäre in diesem Sinne nicht valide, wenn er z. B. den Kompetenzbereich des Leseverstehens nicht berücksichtigen würde. Konstruktirrelevante Varianz liegt hingegen vor, wenn bestimmte Eigenschaften des diagnostischen Verfahrens oder der Zielpopulation, die nichts mit der zu messenden Fähigkeit zu tun haben, eine Aufgabe für bestimmte Probanden(-Gruppen) leichter oder schwerer machen (z. B. die kann die sprachliche Darstellung von mathematischen Anwendungsaufgaben durch die Verwendung von Fremdwörtern ggf. Personen mit weniger elaborierten sprachlichen Kompetenzen bei der Lösung der Mathematikaufgaben benachteiligen).

Am Anfang einer Kompetenzmodellierung wird meist ein inhaltlicher Rahmen erarbeitet, worauf sich die zu beschreibende Kompetenz bezieht. Die Breite und das Auflösungsvermögen (im Sinne von „Granularität“, vgl. Rupp & Mislevy, 2007) dieses Rahmens kann erheblich variieren. Durchaus verschiedenartig sind aber auch die konzeptionellen Grundlagen, auf deren Basis solche Rahmensetzungen vorgenommen werden. Diese reichen von theoretisch fundierten Setzungen, worauf sich das jeweilige Kompetenzkonstrukt bezieht, über normativ orientierten Setzungen – insbesondere auf der Grundlage curricularer Dokumente – bis hin zu eher pragmatischen Setzungen anhand des Aufgabenzuschnitts bestimmter Tätigkeiten in Praxisdomänen (vgl. Abs, 2007; Schaper, 2009).

Bei normativ geprägten, breiten Rahmenkonzepten wird die inhaltliche Validität von Kompetenzmodellierungen in der Regel durch eine konsensuelle

Verständigung unter Expertinnen und Experten erarbeitet und abgesichert. Von mittlerer Breite sind curriculare Rahmenkonzepte, die intendierte Lernergebnisse über ein Schuljahr oder über einige Wochen beschreiben. Diesen Weg verfolgt beispielsweise die „curriculum based evaluation" (CBE), welche empiriegestützte Leistungsmessung zur formativen Unterrichtsentwicklung heranzieht (Howell & Nolet, 2000). Die inhaltliche Validität der verwendeten Testverfahren bzw. -formate begründet sich hierbei in der engen Anbindung an ein konkretes Curriculum. Eine noch geringere Breite bzw. engere Fokussierung weisen Kompetenzmodellierungen auf, die sich auf abgrenzbare Aktivitäten bzw. Tätigkeiten beziehen, wie z. B. die schriftliche Subtraktion im Kontext mathematischer Kompetenzen (Lee & Corter, 2011). Als Ausgangspunkt können dabei etwa fachbezogene kognitive Theorien über mentale Modelle oder Erkenntnisse über bereichsspezifische Lösungsstrategien dienen.

3.2 Kognitive Validierung (substanzielle Validität)

Der substanzielle Aspekt von Validität erweitert die Analyse der inhaltlichen Anforderungen um zwei Punkte. Zum einen wird gefordert, dass nicht nur die Inhalte eines diagnostischen Verfahrens repräsentativ sind, sondern auch die Prozesse, die zur Lösung kognitiver Testaufgaben benötigt werden (Messick, 1989). Darüber hinaus wird gefordert, dass es empirische Belege für das Anwenden dieser Prozesse in der konkreten Situation der Leistungsmessung gibt.

Durch die Konkretisierung der Theorieelemente anhand von Aufgabensituationen wird eine Brücke zwischen theoretischen Kompetenzmodellen und ihrer empirischen Erfassung gebildet. Dieser Prozess der Operationalisierung weist beträchtliche Herausforderungen auf (z. B. in der Selektion der Inhalte aus einem breiten Curriculum sowie bei der Auswahl der Itemkontexte, der Itemtypen und der Erfassungssituation) und lässt sich in der Regel nicht mithilfe algorithmisierter Prozeduren bewerkstelligen. Eine mangelnde Passung zwischen den theoretischen Konstrukten und den tatsächlich ablaufenden Kognitionen, die durch die spezifische Operationalisierung angeregt werden, kann eine ernste Bedrohung der Validität darstellen. Der Weg von allgemeinen Situationen, welche als konstitutiv für einen Kompetenzbereich angesehen werden, bis hin zu den konkreten Erfassungssituationen gleicht einem mehrschrittigen Übersetzungsvorgang, bei dem jedes Mal die inhaltliche Bedeutung beeinträchtigt werden kann. Je nach diagnostischem Format können bspw. unterschiedliche kognitive Bearbeitungsformen ein und derselben Aufgabe angeregt werden. So kann in einem diagnostischen Interview oder Prüfungsgespräch sich ein Schüler oder eine Schülerin aufgrund von Feedback durch den Interviewer korrigieren und möglicherweise mehrere Lösungswege eruieren und den angemessensten wählen. Bei der Bearbeitung

derselben Aufgabe dargereicht als schriftliche Testaufgabe mit offenem Format wird dieselbe Person sich möglicherweise nur auf die Erarbeitung einer Lösung konzentrieren und vermutlich das Ergebnis nicht mehr an der Realsituation validieren. Wird die Aufgabe schließlich drittens in Form einer geschlossenen Multiple-Choice-Aufgabe gestellt, führt dies möglicherweise dazu, dass die Person die Aufgabe so bearbeitet, indem sie eine Entscheidung durch Ausschließen vorgegebener Ergebnisse aufgrund von Plausibilitätsüberlegungen trifft.

Generell kann man zwischen eher proximalen oder eher distalen Operationalisierungen von Kompetenzen unterscheiden: Die Kompetenz „Schulleistung" wird besonders valide durch Aufgabenformate erfasst, wie sie in der Schule tatsächlich üblich sind. Wählt man als Kompetenzbereich „Leistungen in Abschlussprüfungen" (Büchter & Pallack, 2012), verschwindet die Differenz zwischen Konstrukt und Operationalisierung sogar völlig. Shavelson (2010) sieht in einer situationsnahen Erfassung eine besondere Qualität des Kompetenzkonzeptes, welche aber aus testökonomischen Gründen oft nur beschränkt realisiert wird.

Zur Gewährleistung der kognitiven Validität können sowohl theoretisch als auch empirisch fundierte Argumente herangezogen werden. Ein Verfahren, um die kognitive Validität einer Kompetenzmodellierung bereits in der Phase der Modellkonstruktion zu gewährleisten, ist die von Crandall, Klein und Hoffman (2006) beschriebene cognitive task analysis (CTA). Hierbei handelt es sich um unterschiedliche Analysezugänge, die dazu dienen, die kognitiven, aber auch manuellen Tätigkeiten, Abläufe und das jeweilige Wissen und Denken, das bei der Aufgabendurchführung benötigt wird, zu ermitteln. Ein solches Vorgehen kommt der Modellierung von Kompetenzen besonders entgegen, da diese sich ja qua Definition über typische Anforderungssituationen definieren (Weinert, 2001).

Die Überprüfung der kognitiven Validität bereits generierter Testitems kann darüber hinaus mithilfe der Beurteilungen von Expertinnen und Experten, also von solchen Personen, die umfangreiches und vertieftes Wissen über kognitive Prozesse bei der Aufgabenbearbeitung durch eigene Forschung oder Praxis besitzen, generiert werden (Rubio et al., 2003). Entsprechende Nachweise können aber auch durch eine Untersuchung der Bearbeitungsprozesse unter testnahen Bedingungen erzielt werden. Das geschieht in so genannten „cognitive labs" (Snow & Lohman 1989) mit Hilfe der Methode des Lauten Denkens oder stimulated-recall-Techniken (Ericsson & Simon 1993) oder mit der Aufzeichnung und Analyse von Augenbewegungen (z. B. Cohors-Fresenborg et al., 2003). Leighton und Gokiert (2005) nutzten Laute-Denk-Protokolle während der Aufgabenlösung und in einer retrospektiven Form, um die kognitive Validität von Items zur Messung von Logikkompetenzen bei Studierenden eines Einführungskurses in Wissenschaftstheorie zu

überprüfen. Durch die Analysen konnten grundlegende Schwierigkeiten der Studierenden bei der Lösung der Logikaufgaben aufgedeckt werden (z. B. bei der Analyse der Aufgabenanforderungen und der Entwicklung eines mentalen Modells), die zu Anpassungen der Aufgabeninstruktion genutzt werden können.

3.3 Strukturelle Validierung

Der strukturelle Aspekt von Validität nach Messick bezieht sich darauf, ob das bei einer Messung explizit oder implizit zugrunde liegende Messmodell mit den Strukturen des Konstrukts übereinstimmt. Z. B. sollten die Verrechnungsprozeduren, die dazu führen, dass Bewertungen von Qualitätsindikatoren zu einem Gesamtscore zusammengefasst werden, auf Wissen darüber beruhen, wie die an diesen Verhaltensweisen beteiligten Prozesse in ihrem Zusammenwirken den festgestellten Effekt produzieren.

Messick (1995) versteht unter dem strukturellen Aspekt von Validität die Passung des „scoring models" zu den Strukturen des zu erfassenden theoretischen Konstrukts. Das scoring model definiert den Übergang von den Situationen bzw. Aufgaben zu einer zahlenmäßigen Repräsentation des hierbei auftretenden Verhaltens bzw. der auftretenden Lösungen (im Sinne eines Messprozesses). Im Rahmen der klassischen Testtheorie wäre hier also zu überprüfen, ob die Anzahl der Variablen (also die Dimensionalität des Messmodells) und die Bewertung und Gewichtung der Lösungen strukturell valide sind, also eine Passung zur theoretischen Struktur des zu messenden Konstruktes aufweisen. Dies wird meist dadurch geprüft, dass konkurrierende Messmodelle mit unterschiedlichen strukturellen bzw. dimensionalen Annahmen hinsichtlich ihrer Passung bzw. ihres „Fits" zu einem vorhandenen Datensatz und dessen empirischen Zusammenhängen verglichen werden; d. h. das Strukturmodell, dass den besten Fit bzw. die besten Fitkennwerte aufweist wird als valider als die Vergleichsmodelle bzw. als validestes Modell bezeichnet (Embretson & Reise, 2000).

Wenn Kompetenzmodellierung das Paradigma klassischer Testtheorie verlässt und sich probabilistischen Messmodellen mit manifesten und latenten Variablen zuwendet, muss man die Prüfung der strukturellen Validität auf die Charakteristika dieser Art der Messung ausdehnen. Die Frage der strukturellen Validität bezieht sich dann nicht mehr nur auf das scoring model, sondern auch auf das gewählte probabilistische Messmodell.

Psychometrische Ansätze der Kompetenzmodellierung und -messung grenzen sich von klassischen Konzepten der differentiellen Psychologie (z. B. der Intelligenzmessung) durch die Art und Weise ab, in der sie Anforderungen der Situationen einerseits und Eigenschaften von Individuen andererseits betrachten und aufeinander beziehen (McClelland, 1973; Hartig, 2008). Damit stehen solche Ansätze der Kompetenzforschung der Expertiseforschung

und der fachdidaktischen Forschung näher als der Intelligenzforschung. Messmodelle, die so etwas leisten, beschreiben das Verhalten (Response) von Individuen bei bestimmten Situationen (Aufgaben, Items) in probabilistischer Abhängigkeit von Anforderungsmerkmalen der Situationen (Itemmerkmale) und Dispositionen der Personen (latente Personenmerkmale). Sie werden auch als „Probabilistische Testmodelle“ oder „IRT-(Item-Response-Theory)-Modelle“ bezeichnet. Ein solches probabilistisches Messmodell erlaubt die Messung von Aufgabenschwierigkeiten und Personenfähigkeiten auch bei nicht deterministischem Personenverhalten und wird daher grundsätzlich als strukturell passend zu den Grundannahmen der Kompetenzmodellierung angesehen: Eine latente Fähigkeit einer Person führt nicht immer zu demselben Verhalten, wohl aber zu einer Verhaltenstendenz, die durch eine situations- und personenspezifische Erfolgswahrscheinlichkeit beschrieben werden kann.

Das in diesem Zusammenhang beschriebene Rasch-Modell ist aber strukturell so stark vereinfachend, dass man plausibler weise kaum annehmen wird, dass es jegliche Art von Kompetenzen strukturell adäquat abbildet. Das ist auch nicht erforderlich, denn mittlerweile ist dieses Urmodell probabilistischer Kompetenzmessung um eine Vielzahl von Modellvarianten erweitert worden. Dabei werden prinzipiell drei unterschiedliche Typen von Modifikation unterschieden (Rost, 2004; DiBello et al., 2007):

1) Hinzunahme weiterer zu schätzender Parameter, die eine bessere Anpassung an die Daten ermöglichen (z. B. Rateparameter, die rein zufällig richtige Multiple-Choice-Antworten mitberücksichtigen, oder Trennschärfeparameter, die beschreiben, dass Items zwischen Personen hoher und niedriger Kompetenz unterschiedliche gut diskriminieren).
2) Erweiterung auf weitere (insbesondere auch kategoriale) Variablen (z. B. latente Variablen, die das Vorliegen oder Nichtvorliegen bestimmter Teilkompetenzen beschreiben oder die die Zugehörigkeit von Personen oder Aufgaben zu bestimmen Gruppen modellieren; im Sinne von latent-class-Modellen).
3) Strukturelle Modifikationen, z. B. hinsichtlich der Dimensionalität (z. B. als multidimensionales Latent-Variable-Modell) oder hinsichtlich des logischen Zusammenspiels der latenten Fähigkeiten bei der Aufgabenlösung (z. B. als hierarchisches Latent-Variable-Modell).

Mittlerweile stehen verschiedenste Modelle mit einer kaum überschaubaren strukturellen Vielfalt hinsichtlich Skalenniveau sowie Anzahl und strukturellem Zusammenspiel der Variablen zur Verfügung. Die Prüfung der strukturellen Validität bei der Wahl des Messmodells wird dadurch deutlich komplexer bzw. differenzierter; es geht nicht mehr nur um die Frage der Dimensionalität. Vielmehr hat man die Möglichkeit, aus der Vielfalt

der Modelle die auszuwählen, die das Verhalten der Probanden auf eine möglichst passende Weise beschreiben und damit eine reliable und valide Messung der zu erfassenden Kompetenzen ermöglichen. Wenn eine solche Passung jedoch nicht exploratorisch durch Anpassung hinreichend vieler Parameter geschehen soll, braucht es a priori eine strukturelle Korrespondenz zwischen gewähltem psychometrischem Modell und dem zu modellierenden Kompetenzbereich (Rupp & Mislevy, 2007).

Hartig (2008) zählt einige wesentliche Kriterien für eine strukturell valide Modellwahl (also theoretische Validitätsargumente im Messick'schen Sinne) auf, welche sich auf die Passung zwischen theoretischem und psychometrischem Modell beziehen:

- Sind die latenten Personenvariablen eher als kontinuierlich (Skalen) oder als kategorial (Typen) anzunehmen?
- Wie viele unabhängig anzunehmende Teilkompetenzen sollen das Verhalten beschreiben, d. h. wie viele Dimensionen sind plausibel?
- Erfordern die Aufgaben jeweils nur eine Teilkompetenz (between-item dimensionality) oder müssen bei einigen Aufgaben Kompetenzen aus mehreren Dimensionen zusammenkommen (within-item dimensionality)?
- Können sich die Teilkompetenzen bei einer Aufgabe gegenseitig ersetzen oder ergänzen (kompensatorische Modelle) oder werden mehrere Teilkompetenzen zugleich benötigt, um die Aufgabe erfolgreich zu bewältigen (nicht-kompensatorische Modelle)?

Die strukturelle Passung zwischen theoretischem Kompetenzmodell und psychometrischem Messmodell ist also nicht primär eine Frage der (An)Passung eines vielparametrigen mathematischen Modells an gegebene Daten, sondern eine theoriegeleitete Entscheidung, die theoretisch wie empirisch auf ihre Validität geprüft werden kann. Ein theoretisches Argument für die strukturelle Validität kann darin bestehen, die Skalenqualität oder das Zusammenspiel der latenten Variablen mit dem Wissen um die kognitiven Prozesse bei der Aufgabenlösung zu begründen: Ist es beispielsweise plausibler anzunehmen, ein Kind habe entweder die Fähigkeit, bei der schriftlichen Addition Überträge zu berücksichtigen oder nicht? Oder gibt es diese Fähigkeit in graduellen Abstufungen? Ein empirisches Argument für die strukturelle Validität kann durch den Vergleich konkurrierender Modelle auf ihre Passung zu empirischen Daten gewonnen werden (siehe oben).

3.4 Verallgemeinerungsbezogene Validierung

Die inhaltliche Repräsentativität eines Tests gewährleistet, dass sich die Interpretation der Testergebnisse nicht nur auf die im Test enthaltenen Aufgaben bezieht, sondern auf das Konstrukt im Ganzen verallgemeinern lässt. Ein Ergebnis in z. B. einem inhaltsvaliden Sprachtest sollte also eine Aussage

über die (fremd-)sprachliche Kompetenz im Allgemeinen zulassen. Ein typisches Problem bei einer zeitgebundenen Testung ist immer die Abwägung zwischen der Breite, in der man ein Konstrukt abdeckt, und der Tiefe (Genauigkeit). Für die Generalisierbarkeit der Ergebnisse ist eine Repräsentativität der Inhaltsbereiche, also eine breite Abdeckung, von Bedeutung. Der Aspekt der Generalisierbarkeit behandelt darüber hinaus Verallgemeinerungen der Testergebnisse über verschiedene Zeitpunkte, Situationen und Beurteiler.

Das Ziel einer Kompetenzmodellierung ist die Konstruktion eines möglichst allgemeingültigen Messmodells für einen bestimmten Kompetenzbereich, d. h. eines Instruments, welches eine Aussage treffen kann, die nicht von bestimmten Bedingungen des Konstruktionsprozesses abhängt. Zu diesen Bedingungen zählen unter anderem die Festlegung der Stichprobe, die Auswahl der Items aus einem möglichen Itemuniversum oder das Verfahren der Antwortbewertung (etwa durch Beurteiler). Die so genannte Generalisierbarkeitstheorie (Webb, Shavelson & Haertel, 2007) ermöglicht im Paradigma der klassischen Testtheorie eine Quantifizierung solcher unerwünschter Varianz und liefert so empirische Argumente für die Verallgemeinerbarkeit eines Konstruktes.

Im Zusammenhang mit der Verallgemeinerbarkeit wird oft die Stichprobenunabhängigkeit der Parameterschätzung bei IRT-Modellen genannt (z. B. bei Cavanagh, 2011, S. 112): Wenn das gewählte Modell sich empirisch an einem Datensatz bestätigt, so ergeben sich dieselben Werte für Personenfähigkeiten und Aufgabenschwierigkeiten, auch wenn man nur eine Teilmenge der Personen oder Items zur Schätzung heranzieht. Diese als „spezifische Objektivität“ bezeichnete Eigenschaft erweist sich als Argument für die Fairness des Instruments und hat viele praktische Implikationen, wie z. B. die Reduktion von Testbelastungen oder die Erhöhung von Messgenauigkeit durch Multimatrixdesigns oder computeradaptives Testen (Frey, 2007). Es handelt es sich aber gewissermaßen nur um ein Argument für die Verallgemeinerbarkeit innerhalb der Modellierung, das nicht herangezogen werden kann, wenn man über die konkreten Items des Tests oder die Personen der Normierungsstichprobe hinausgeht. Auch die Analyse der Unabhängigkeit der Modellierung von speziellen Untergruppen (also z. B. Geschlecht oder Herkunft), wie sie bei einer IRT-Modellierung beispielsweise durch differential item functioning (DIF)-Analysen umgesetzt werden kann (Holland & Wainer, 1993), ist ein solches empirisches Argument für die Verallgemeinerbarkeit innerhalb der Erhebungspopulation. Es gibt allerdings keine Garantie dafür, dass ein Kompetenzmodell auch bei einer neuen Stichprobe oder bei einem Retest nach einer längeren Lernepisode strukturell stabil bleibt und sich daher für einen Gruppenvergleich oder eine Längsschnittmessung eignet. In der Frage der so genannten Robustheit von Kompetenzmodellen steht die Forschung daher noch am Anfang (vgl. Robitzsch, 2013).

3.5 Externe Validität

Der Aspekt der externen Validität bezieht sich vor allem auf hypothesenkonforme Zusammenhänge der Testergebnisse mit Außenkriterien. Die Zusammenhänge zwischen der Messung und externen Kriterien sollte den theoretisch erwarteten Zusammenhängen entsprechen (Kane, 2006). Dieser Validierungsaspekt entspricht somit der „klassischen" Kriteriumsvalidität, unter der die Vorhersage einer direkt beobachtbaren Verhaltensweise außerhalb der Testsituation – daher auch externe Validität – als Kriterium für die Gültigkeit eines Diagnoseverfahrens verstanden wird (Blickle, 2014). 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).

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 externale bzw. 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. Der Grad, mit dem Kriterien aufgrund eines Testergebnisses prädiziert werden können, hängt dabei allerdings von der Objektivität und Reliabilität der Messung des Prädiktors sowie von der Objektivität, Reliabilität sowie Inhalts- und Konstruktvalidität des Kriteriums ab (Blickle, 2014).

Typische Probleme mit der Objektivität einer Kriteriumsmessung sind z. B. subjektive Urteilstendenzen (z. B. Halo-Effekte) von Fremdbeurteilungen.

Bezüglich der Reliabilität eines Kriteriums kann darüber hinaus problematisch sein, 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.

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 (Marcus & Schuler, 2006). Ist der Merkmalsbereich durch das Kriterium nicht vollständig abgedeckt, spricht man von Kriteriumsdefizienz (a.a.O.). 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 (a.a.O.).

Speziell im Bereich der Kompetenzforschung, stellt sich als zusätzliche Herausforderung, dass zwar angenommen wird, Kompetenz unterliege als latente Disposition erfolgreichem Handeln in Realsituationen (Performanz), dass zwischen diesen beiden Merkmalen aber zahlreiche vermittelnde Prozesse stattfinden, die einen direkten Nachweis externaler Validität schwierig machen (Schaper, 2013). Wass et al. (2001) und Albino et al. (2008) haben daher einen vierschrittigen Validierungsprozess vorgeschlagen, der sequentiell Zwischenschritte wie kognitive Umstrukturierungen und unterschiedliche Rahmenbedingungen berücksichtigt und damit erfolgsversprechender zu sein scheint. Dieser Ansatz baut auf dem Konzept der Kompetenzpyramide nach Miller (1990) auf. Dieser Autor hat, aufbauend auf Taxonomien kognitiver Prozesse, den Grad der Authentizität der Kompetenzprüfung methodologisch zwischen vier Assessment-Formaten unterschieden, mit denen professionelle Kompetenz (hier speziell bei 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 Bedingungen des beruflichen Alltags demonstriert werden muss (does).

Diese vier Assessment-Formate können methodisch wie folgt umgesetzt 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, oder mithilfe schriftlicher bzw. mündlicher Erhebungsformate geschehen.
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 (siehe z. B. 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). Dabei wird eine holistische Betrachtung dessen vorgenommen, was gekonnt wird, beispielsweise dokumentiert über Videoaufnahmen und daran anschließende standardisierte Kodierungen und Bewertungen der Aufzeichnungen (siehe z. B. Vogelsang, 2014).

Von externer Validität spricht man im Messick'schen Verständnis außerdem dann, wenn ein Konstrukt in eine systematische theoretische und empirische Beziehung mit bestehenden Theorien und anderen Konstrukten gestellt wird (Messick, 1995). Die Überprüfung der so genannten konvergenten und diskriminanten Validität von Kompetenzmodellierungen ist Ausdruck des Bemühens um die Einbindung einer modellierten Kompetenz in ein so

genanntes nomologisches Netz (Cronbach & Meehl, 1955). Konvergente Validität bezeichnet in diesem Zusammenhang den Grad, in dem ein Konstrukt von verschiedenen Verfahren übereinstimmend (konvergent) gemessen wird (z. B. durch zwei unterschiedliche Tests zur Erfassung des erziehungswissenschaftlichen Wissens in der Lehrerausbildung, vgl. Seifert & König, 2012). D. h. ein Konstrukt muss mit anderen Verfahren, welche auch dieses Konstrukt erfassen, ähnlich gemessen werden können wie mit dem zu validierenden Verfahren. Die konvergente Validität ergibt sich aus der Korrelation des Zielkonstrukts mit demselben Konstrukt anderer Verfahren, wobei sie somit möglichst hoch ausfallen sollte. Diskriminante Validität bezeichnet hingegen den Grad, in dem ein Verfahren zwischen verschiedenen Konstrukten diskriminiert, also unterscheidet (z. B. die Korrelation zwischen Intelligenz und Gewissenhaftigkeit). Das Kriterium der diskriminanten Validität fordert, dass sich das Zielkonstrukt von anderen Konstrukten unterscheidet (z. B. zwischen Konzentrationsfähigkeit und Intelligenz). Die diskriminante Validität berechnet sich aus der Korrelation zwischen dem Zielkonstrukt und anderen sich theoretisch deutlich unterscheidenden Konstrukten, wobei die entsprechenden Zusammenhänge also möglichst gering sein sollten. Wenn die diskriminante Validität zu hoch ist, ist das Zielkonstrukt daher nicht genügend von anderen Konstrukten abgegrenzt, woraus z. B. geschlossen werden kann, dass die Items des Tests im Sinne einer besseren Abgrenzung der diskriminanten Konstrukte überarbeitet werden sollten. Bei der Einordnung und Überprüfung eines Kompetenz-Konstrukts in ein nomologisches Netz werden somit Annahmen darüber formuliert, mit welchen anderen Variablen das zu erfassende Konstrukt in welchem Zusammenhang stehen sollte (Cronbach & Meehl, 1955). Das entsprechende Netzwerk 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 hierzu auch das Vorgehen bei der Multitrait-Multimethod-Methode zum Nachweis von konvergenter und diskriminanter Validität nach Campbell und Fiske,1959). Entspricht das Zusammenhangsmuster den theoretisch erwarteten Zusammenhängen, unterstützt dies sowohl die Interpretation der Testwerte bezogen auf das (Kompetenz-) 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).

In diesem Sinne sollte im Kontext der Kompetenzforschung angestrebt werden, die bereits entwickelten Kompetenzmodellierungen zu vernetzen, und nicht etwa isolierte Kompetenzmodelle für immer neue Kompetenzfacetten zu entwickeln und nebeneinander zu stellen. Dabei könnte es z. B. um solche Frage gehen, inwieweit Kompetenzdimensionen aus großen Survey-Studien mit Kompetenzstrukturmodellen aus enger fokussierenden Studien zusammenhängen, oder welche Beziehungen zwischen Kompetenzmodellen für fachliche und überfachliche Konzeptualisierungen verwandter Kompetenzen bestehen (also z. B. zwischen mathematischem Problemlösen und Metakognition oder figuralem Denken)?

3.6 Konsequentielle Validierung

Der Aspekt der „consequential validity" bezieht sich darauf, ob die angestrebten Effekte der durchgeführten Diagnose eingetreten und ob nicht-intendierte Wirkungen ausgeblieben sind, und dies sowohl kurz- als auch langfristig. Im Rahmen des Messick'schen Validitätsansatzes nimmt dieser Aspekt eine große Bedeutung ein. Validität ist in diesem Zusammenhang immer auch die empirische Evidenz dafür, dass die Interpretation der Ergebnisse und die daraus abgeleiteten Konsequenzen angemessen sind.

Da Kompetenzmodelle und -messinstrumente nicht nur im Bereich der Grundlagenforschung eingesetzt werden, sondern zunehmend auch zur Erhebung und Rückmeldung von Leistungen im Schulsystem, lässt die von Messick (1995) angemahnte argumentative und empirische Prüfung ihrer jeweils in Anspruch genommenen Ziele und Wirkungen bedeutender denn je erscheinen. In Deutschland werden Kompetenzmodelle und -messinstrumente zurzeit vor allem im Kontext folgender Bereiche eingesetzt bzw. genutzt (vgl. Helmke & Hosenfeld, 2005; Klieme, Hartig & Rauch, 2008):

- Bildungsmonitoring auf Systemebene
- Rückmeldung von Leistungsdaten auf Klassen- und Schulebene mit dem Ziel der Schul- und Unterrichtsentwicklung
- Professionalisierung im Sinne der Förderung diagnostischer Kompetenzen von Lehrkräften
- Unterstützung individueller Diagnose zur Vorbereitung pädagogischer Förderentscheidungen

Eine Bedrohung der konsequentiellen Validität wird z. B. darin gesehen, dass die Zielsetzung eines Verfahrens oftmals zu breit definiert bzw. angenommenen werden:

> „Often a single assessment is used for multiple purposes; in general, however, the more purposes a single assessment aims to serve, the more each purpose will be compromised. For instance, many state tests are used for both individual and program assessment purposes. This is not necessarily a problem, as long as assessment designers and users recognize the compromises and trade-offs such use entails“ (Pellegrino et al., 2001, S. 161).

Zur konsequentiellen Validität einer Kompetenzmodellierung zählt mithin die Passung des gewählten Kompetenzmodells und der mit ihm möglichen Aussagen zu den intendierten Nutzungsweisen. Diese Passung kann bereits bei der Modellkonstruktion berücksichtigt werden (vgl. Hartig, 2008; Rupp & Mislevy, 2007).

Das wohl am weitesten in den Unterricht reichende Format von Kompetenzmodellierungen findet man bei den unterschiedlichen Formen von zentralen Lernstandserhebungen bzw. Vergleichsarbeiten (Helmke & Hosenfeld, 2005). Auch wenn diese nicht die weitreichenden Konsequenzen fordern, die mit analogen Testungen in den Vereinigten Staaten verbunden sind, so sind die erwünschten und unerwünschten Rückwirkungen auf Schule und Unterricht doch in der Diskussion (z. B. Altrichter, 2010). Mit der Kompetenzmodellierung verbunden werden bisweilen recht weitreichende Annahmen über die Möglichkeiten einer darauf aufbauenden Unterrichtsentwicklung (z. B. Peek & Dobbelstein, 2006). Entsprechende Analysen zeigen allerdings, dass die Wirkungen solcher zentraler Kompetenzerhebungen eher gering bis zweifelhaft sind. So erfassten beispielsweise Wacker und Kramer (2012) die Einschätzungen von Lehrkräften in Baden-Württemberg ($n>700$) zu den wahrgenommenen Funktionen zentraler Lernstandserhebungen jeweils zu Beginn und vier Jahre nach ihrer Einführung. Sie fanden einen signifikanten Wechsel von einer anfänglichen Zustimmung zu einer Ablehnung der Aussage, Lernstandserhebungen böten einen Orientierungsrahmen zur Unterrichtsplanung, zur Leistungsbeurteilung und zum Erkennen von Lernrückständen.

Ein möglicher Grund für eine entsprechend geringe konsequentielle Validität dieser Ansätze kann in der mangelnden Passung der Kompetenzmodellierung und insbesondere der Rückmeldeformate zu den Bedürfnissen und Kompetenzen der Lehrkräfte liegen: Die zum Zwecke des Klassenvergleichs in der Regel eindimensional angelegten Kompetenzmodelle werden bei der Rückmeldung zwar durch inhaltlich beschriebene Kompetenzstufen ergänzt. Lehrkräfte erhalten aber hierdurch oft keine hinreichend spezifischen Impulse für ihre Unterrichtsgestaltung.

Der Aspekt der konsequentiellen Validität ist – darauf sei abschließend hingewiesen – der umstrittenste – nicht nur, weil dieser schwierig empirisch zu prüfen ist und sich nicht gut in das Rational der Konstruktvalidität

einpasst. Die konsequentielle Validität nimmt die Testentwicklung auch ein Stück weit in Haftung für den späteren Einsatz ihrer Tests. Man will damit zumindest erreichen (vgl. Kane, 2013), dass eine Unterstützung des Testeinsatzes in der Praxis durch die ursprünglichen Testentwickler gefordert wird – ggf. auch noch lange nachdem diese das Projekt beendet haben. Die Betonung dieses Validitätsaspekts lässt sich möglicherweise vor dem Hintergrund der zahlreichen negativen Erfahrungen erklären, die in den USA mit Testprogrammen gemacht wurden (Blömeke, 2013). Eine Reihe von Autoren (z. B. Borsboom et al., 2004; Scriven, 2010) widersprechen daher insbesondere dem Einbezug konsequentieller Validitätsaspekte, da die Forderung, Konsequenzen eines Testeinsatzes in die Validierung einschließen zu müssen, manche Validierungen praktisch undurchführbar mache. Zudem könne eine missbräuchliche Verwendung niemals ausgeschlossen werden.

4 Fazit

Als Fazit kann festgehalten werden, dass es im Zusammenhang mit der Kompetenzmodellierung und -messung angemessener erscheint, nicht von der Validität eines Kompetenztests zu sprechen, sondern jeweils die Validität verschiedener Interpretationen von Testergebnissen zu betrachten. Für eine Validierung eines Kompetenzmodells und des entsprechenden Tests gilt es daher zunächst zu spezifizieren, auf welche Interpretationen eines Testergebnisses sich die intendierte Validität bezieht.

Generell ist in diesem Zusammenhang festzustellen, dass Testergebnisse bzw. auch individuelle Testwerte in vielfältiger Weise interpretiert werden können, was in Bezug auf die Qualitätssicherung von Tests dazu führt, dass auch die möglichen Strategien zur Validierung entsprechender Interpretationen vielfältig sind. Bei der Entwicklung eines Kompetenztests steht man daher vor der Frage, wie man mit diesen vielfältigen Anforderungen der Validierung umgehen soll; denn für jede mögliche Interpretation und Verwendung eines Testergebnisses Argumente oder empirische Belege zu erbringen, ist ein sicherlich wünschenswertes, aber oftmals unrealistisches Unterfangen. In der Regel lässt sich jedoch relativ leicht entscheiden, welche Interpretationen und Verwendung für einen Test besonders relevant und zentral sind. Hieraus lassen sich Prioritäten ableiten, welche Validierungstrategien am dringlichsten verfolgt werden sollten (vgl. Hartig et al., 2008). Bei der Entwicklung und dem Einsatz eines Kompetenztests sollte daher sichergestellt sein, dass die jeweils wichtigsten Interpretationen der Testergebnisse empirisch gestützt sind. Kritisch und vorsorglich sollte aber auch diskutiert werden, welche Interpretationen der Testergebnisse nahe liegend oder wünschenswert sind, zum gegenwärtigen Zeitpunkt aber

noch nicht als gestützt durch entsprechende Evidenzen betrachtet werden können.

In vielen Beiträgen dieses Bandes kommt ein solches interpretationsspezifisches Vorgehen zur Validierung von Testinstrumenten zum Ausdruck. In unterschiedlichen Konstellationen werden mindestens vier der sechs Aspekte des umfassenden Validitätsverständnisses nach Messick (1995) konzeptionell dargelegt oder geprüft. Diese Beispiele eignen sich darum als abschließende, detaillierte Illustrationen der hier vorgenommenen Auseinandersetzung mit dem Validitätsbegriff und können zum gezielten Lesen der entsprechenden Beiträge anregen.

Brückner, Zlatkin-Troitschanskaia & Förster (in diesem Band) berufen sich auf das hier vorgestellte Validitätsverständnis und betonen, dass sich die Sicherstellung der Validität auf den gesamten Prozess der Testentwicklung beziehen muss. In ihrem KoKoHs Projekt (WiWiKom) steht dabei das Ziel einer internationalen Vergleichbarkeit von Testwertinterpretationen im Mittelpunkt. Die Gültigkeit international vergleichender Testwertinterpretationen wurde in sechs Abschnitten der Instrumentenentwicklung mit jeweils spezifischen Strategien sichergestellt, die vier der sechs Aspekte des Messick'schen Validitätsverständnisses zugeordnet werden können. So konzentrieren sich die Abschnitte „domain analysis" und „assessment implementation" auf die inhaltliche Validierung und die Abschnitte „domain modeling", „Assessment frameworks" sowie „assement implementation" schwerpunktmäßig auf die kognitive Validierung. Der Abschnitt „assessment delivery" deckt die strukturellen und verallgemeinerungsbezogen Validitätsaspekte ab.

Der Beitrag von Ştefănică, Behrendt, Dammann, Nickolaus & Heinze (in diesem Band) konzentriert sich mit Analysen der Modulhandbücher auf die Inhaltsvalidierung und leitet aus aktuellen Kompetenzmodellierungen in beruflichen Domänen kognitive Strukturen ab, die einer empirischen Prüfung zugänglich sind. Ähnlich verfahren Musekamp & Spöttl (in diesem Band).

Bei Musekamp, Schlömer & Mehrafza (in diesem Band) wird die kognitive Validität des Tests über die Vorhersage von Itemschwierigkeiten mittels theoretisch hergeleiteter Itemmerkmale geprüft. Saniter (in diesem Band) hingegen wählt einen anderen Weg und untersucht die empirischen Antworten von Studierenden auf offene Testaufgaben, um aus den gewählten Lösungsansätze Rückschlüsse auf die zugrunde liegenden Denkprozesse zu ziehen. Beide Ansätze der kognitiven Validierung sind insbesondere für Testwertinterpretationen geeignet, die individuelle Rückmeldungen über Stärken und Schwächen von Studierenden erlauben sollen.

Literatur

Abs, H. J. (2007). Überlegungen zur Modellierung diagnostischer Kompetenz bei Lehrerinnen und Lehrern. Frankfurt: Deutsches Institut für Internationale Pädagogische Forschung.

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.

Altrichter, H. (2010). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In: Herbert Altrichter und Katharina Maag Merki (Hrsg.), Handbuch Neue Steuerung im Schulsystem. Wiesbaden: Verlag für Sozialwissenschaften 2010, S. 219 – 254.

Blickle, G. (2014). Leistungsbeurteilung. In Nerdinger, F., Blickle, G. & Schaper, N. (2011). Lehrbuch Arbeits- und Organisationspsychologie, (S. 271–290). Heidelberg, Berlin, New York: Springer.

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., 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.

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

Borowski, A., Neuhaus, B. J., Tepner, O., Wirth, J. & Fischer, H. et al. (2010). Professionswissen von Lehrkräften in den Naturwissenschaften (ProwiN) – Kurzdarstellung des BMBF-Projekts. Zeitschrift für Didaktik der Naturwissenschaften 16, 341–348.

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. Heidelberg: Springer.

Büchter, A. & Pallack, A. (2012). Zur impliziten Standardsetzung durch zentrale Prüfungen – methodische Überlegungen und empirische Analysen. Journal für Mathematik-Didaktik, 33 (1), 59–85.

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

Cavanagh, R. F. (2011). Establishing The Validity of Rating Scale Instrumentation in Learning Enviornment Investigations. In R. F. Cavanagh and R. F. Waugh (Ed.), Applications of Rasch Measurement in Learning Environments Research, pp. 101–118. Rotterdam, Netherlands: Sense Publishers.

Cohors-Fresenborg, E., Brinkschmidt, S., & Armbrust, S. (2003). Augenbewegungen als Spuren prädikativen oder funktionalen Denkens. Zentralblatt für Didaktik der Mathematik, 35(3), 86–93.

Crandall, B., Klein, G. & Hoffman, R. R. (2006). Working Minds: A Practitioner's Guide To Cognitive Task Analysis. Cambridge, MA: MIT Press.

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

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In: C. R. Rao & Sinharay (Hrsg.). Handbook of Statistics (pp. 979–1030). New York: Elsevier.

Embretson, S. E. & Reise, S. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum Publishers.

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

Frey, A. (2007). Adaptives Testen. In: Moosbrugger, H. & Kelava, A. (Hrsg.), Testtheorie und Fragebogenkonstruktion (S. 261–278). Berlin: Springer.

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

Hartig, J. (2008). Psychometric Models for the Assessment of Competencies. In: Hartig, J., Klieme, E., & Leutner, D. (Hrsg.) Assessment of competencies in educational contexts. Cambridge, Mass. u. a.: Hogrefe.

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. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), Test- und Fragebogenkonstruktion. (S. 143–171). Berlin: Springer.

Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik- Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung, 17–36. Bonn: BMBF.

Helmke, A. & Hosenfeld, I. (2005). Ergebnisorientierte Unterrichtsevaluation. In: Interkantonale Arbeitsgemeinschaft Externe Evaluation von Schulen (Hrsg.), Schlüsselfragen zur externen Schulevaluation (S. 127–151). Bern: h.e.p.-Verlag.

Holland, P. W. & Wainer, H. (1993). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum.

Howell, K. W. & Nolet, V. (2000). Tools for assessment. In Howell, K. W. & Nolet, V. (Eds.), Curriculum-Based Evaluation, Teaching and Decision Making. Scarborough, Ontario: Wadsworth/Thompson Learning.

Jenßen, L., Dunekacke, S. & Blömeke, S. (2014). Qualitätssicherung in der Kompetenzforschung: Standards für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis.

Kane, M. (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(4), 319–342.

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

Kane, M. T. (2013). Validation as a Pragmatic, Scientific Activity. Journal of Educational Measurement, 50(1), 115–122.

Klieme, E., Hartig, J., Rauch, D. (2008). The concept of competence in educational contexts. In Hartig, Johannes et al. (Eds): Assessment of competencies in educational contexts (S. 3–22). Göttingen: Hogrefe.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. Zeitschrift für Pädagogik 52, 876–903.

Seifert, A. & König, J. (2012). Pädagogisches Unterrichtswissen – bildungswissenschaftliches Wissen. Validierung zweier Konstrukte. In J. König & A. Seifert (Hrsg.), Lehramtsstudierende erwerben pädagogisches Professionswissen: Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung (S. 215–233). Münster: Waxmann.

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.

Lee, J., & Corter, J. E. (2011). Diagnosis of Subtraction Bugs Using Bayesian Networks. Applied Psychological Measurement, 35(1), 27–47.

Leighton, J. P. und Gokiert, R. J. (2005). Investigating Test Items Designed to Measure Higher-Order Reasoning using Think-Aloud Methods: Implications for Construct Validity and Alignment. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), Montreal, Quebec, Canada.

Leuders, T. (2014). Modellierungen mathematischer Kompetenzen – Kriterien für eine Validitätsprüfung aus fachdidaktischer Sicht. Journal für Mathematikdidaktik

Lienert, G. & Raatz, U. (1994). Testaufbau und Testanalyse. Weinheim: Beltz.

Marcus, B. & Schuler, H. (2006). Leistungsbeurteilung. In H. Schuler (Hrsg.), Lehrbuch Personalpsychologie (S. 433–470). Göttingen: Hogrefe.

McClelland, D. C. (1973). Testing for competence rather than for intelligence. American Psychologist, 28, 1–14.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (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.

Peek, R. & Dobbelstein, P. (2006). Zielsetzung: Ergebnisorientierte Schul- und Unterrichtsentwicklung. Potenziale und Grenzen der nordrheinwestfälischen Lernstandserhebungen. In Böttcher, W., Holtappels, H. G. & Brohm, M. (Hrsg.), Evaluation im Bildungswesen; Eine Einführung in Grundlagen und Praxisbeispiele (S. 177–194). Weinheim und München, Juventa.

Pellegrino, J; Chudowsky, N., & Glaser, R. (eds) (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Robitzsch, A. (2013). Wie robust sind Struktur- und Niveaumodelle? Wie zeitlich stabil und über Situationen hinweg konstant sind Kompetenzen? Zeitschrift für Erziehungswissenschaft, 16, 41–45.

Rost, J. (2004). Lehrbuch Testtheorie – Testkonstruktion. Bern, Göttingen: Huber.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. Social Work Research, 27, 94–104.

Rupp, A. A. & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Eds.), Cognitive diagnostic assessment in education: Theory and applications (pp. 205–241). Cambridge: Cambridge University Press.

Schaper, N. (2009). Aufgabenfelder und Perspektiven bei der Kompetenzmodellierung und messung in der Lehrerbildung. Lehrerbildung auf dem Prüfstand, 2(1), 166–199.

Schaper, N. (2012). Fachgutachten zur Kompetenzorientierung in Studium und Lehre. Bonn: Hochschulrektorenkonferenz – nexus.

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). Rethinking evaluation methodology. Journal of MultiDisciplinary Evaluation, 6 (13), i–ii.

Shavelson, R. J. (2010). On the measurement of competency. Empirical research in vocational education and training 2(1), 41–63.

Snow, R. E. & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Hrsg.), Educational measurement (S. 263–331). New York: American Council on Education and Mac Millan Publishing Company.

Vogelsang, C. (2014). Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften – Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz. Unveröffentl. Dissertationsschrift, Fakultät für Naturwissenschaften, Universität Paderborn.

Wacker, A. & Kramer, J. (2012). Vergleichsarbeiten in Baden-Württemberg. Zur Einschätzung von Lehrkräften vor und nach der Implementation. Zeitschrift für Erziehungswissenschaften, 15(4), 683–706.

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

Webb, N. M., Shavelson, R. J. & Haertel, E. H. (2007). Reliability and Generalizability Theory. In Rao, C. R. Handbook of Statistics.

Weinert, Franz E. (2001): Concept of Competence: A Conceptual Clarification. In Rychen, D. S. & Salganik, L. (Hrsg.), Defining and Selecting Key Competences (S. 45–65). Seattle: Hogrefe & Huber.

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.