From test validity to construct validity … and back?

Jerry A. Colliver,[1] Melinda J. Conlee[1] & Steven J. Verhulst[2]

CONTEXT Major changes in thinking about validity have occurred during the past century, shifting the focus in thinking from the validity of the test to the validity of test score interpretations. These changes have resulted from the 'new' thinking about validity in which *construct validity* has emerged as the central or unifying idea of validity today. Construct validity was introduced by Cronbach and Meehl in the mid-1950s in an attempt to address the validity of those many psychological concepts that have no clear referent in reality. To do this, construct validity theory required a nomological network – an elaborate theoretical network of constructs and observations connected by scientific laws – to validate the constructs. However, nomological networks are hard to come by and none that would do the job required by construct validity has been forthcoming to date. Thus, the current construct validity approach has retreated to one of simply 'interpretation and argument', but this seems to be too general to tie down the constructs in the way a nomological network would do to give credibility to the validity of the construct. As a result, the concept of validity seems to have been watered down and the credibility of validity claims weakened.

OBJECTIVES The purpose of this paper is to encourage a discussion of the use of construct validity in medical education, and to suggest that test developers and users reconsider the use of abstract theoretical constructs that have no referent apart from theory.

METHODS We present a critical review of these concerns about construct validity and provide for contrast a brief overview of a recently proposed view of measurement based on scientific realism and causality analysis.

*Medical Education 2012: **46**: 366–371*
doi:10.1111/j.1365-2923.2011.04194.x

Read this article online at www.mededuc.com 'read'
Discuss ideas arising from this article at www.mededuc.com 'discuss'

[1]Department of Medical Education, Southern Illinois University School of Medicine, Springfield, Illinois, USA
[2]Department of Statistics and Research Consulting, Southern Illinois University School of Medicine, Springfield, Illinois, USA

Correspondence: Jerry A. Colliver, PhD, Department of Medical Education, Southern Illinois University School of Medicine, 913 North Rutledge Street, Room 2103, Springfield, Illinois 62794-9623, USA. Tel: 00 1 217 545 7765; Fax: 00 1 217 545 4455; E-mail: jcolliver@siumed.edu

INTRODUCTION

The concept of validity has undergone major changes throughout the last century, evolving from an approach that focused on what might be called the fundamental concept of test validity to the current view, construct validity, which has emerged as the central or unifying idea of validity today.[1–4] The focus of thinking has shifted from the validity of the test to the validity of test score interpretations. However, this shift in thinking seems to have weakened the concept of validity and the credibility of validity claims. In this paper, we present a critical review[5] of these changes and discuss the construct validity approach to measurement in terms of its promising rationale and its shortcomings in practice. We then consider a recent view of measurement that proposes a shift of the focus of measurement from theoretical constructs to more reality-based attributes.[6–9] We hope that this discussion will stimulate and clarify thinking about measurement and validity in assessment research and practice in medical education.

THE FUNDAMENTAL CONCEPT OF TEST VALIDITY

The fundamental concept of validity refers to whether a test, or a measurement instrument, measures what it purports to measure. In 1927, Kelly said: '…a test is valid if it measures what it purports to measure.'[10] In 1954, Anastasi wrote: '…validity, i.e. the degree to which the test actually measures what it purports to measure...'[11] To determine whether a test in fact measures what it purports to measure, various methods or approaches have been developed and employed; these were initially referred to as 'validity types'. In the first half of the 20th century, the primary approaches to determining validity were criterion validity and content validity.[1,12] These referred to properties of the *test* itself: that is, whether the *test* provides an accurate estimate of the criterion it purports to measure (current or future) and whether the *test* adequately represents the universe of behaviours it is supposed to measure. Up to the middle of the 20th century, criterion validity (concurrent and predictive) and content validity were *the* validity types – the primary methods used to establish test validity.

CONSTRUCT VALIDITY BASED ON NOMOLOGICAL NETWORKS

Then, in 1954, the American Psychological Association, in its Technical Recommendations for Psychological Tests and Diagnostic Techniques,[13] introduced the idea of construct validity to validate theoretical attributes or qualities that cannot be explicitly defined in terms of a criterion or a universe of behaviours. Cronbach and Meehl were members of the Technical Recommendations Committee (Cronbach was chair) and, in 1955, they published their classic paper, 'Construct validity in psychological tests',[4] which identified validation procedures to obtain evidence relevant to construct validity. This evidence included various aspects of criterion validity and content validity, such that construct validity came to be seen as the unifying concept of validity – not a new 'type' of validity, a third type to be added to criterion validity and content validity – but a conceptual umbrella that covered all thinking about validity, represented a unifying conceptualisation of validity.[2,12] The revolutionary idea underlying Cronbach and Meehl's thinking – which made the unification possible – was that scientific theory testing was seen as part and parcel of test validity, that test validity was determined by theory testing, or 'validation as hypothesis testing' as one author described it.[14]

In construct validity theory, the construct (e.g. intelligence, clinical reasoning, empathy, burnout, professionalism, systems-based practice, etc.) is a postulated or theoretical concept that is defined by its position in a network of other constructs. The relationships among the constructs in the network are defined by scientific laws that link the constructs and form the network. Cronbach and Meehl referred to this as a 'nomological network', which is basically a network of laws that relates constructs: scientific theory.[4] Construct validity, then, is established by any evidence that supports the nomological network of constructs and laws that contains the construct. With the introduction of construct validity, understandings of the concept of validity shifted from the issue of whether a test measures what it purports to measure to the relationship(s) between the construct and other constructs as specified by the nomological network.

Construct validity theory was appealing at the time (in the mid-1950s) because it was consistent with the philosophy of science that dominated scientific psychology, namely, logical positivism.[2,15] Positivists wanted to avoid any reference to 'reality' in scientific theory and criticised the use of theoretical terms (constructs) that claimed to refer to something apart from the theory itself; they saw this practice as metaphysical and thought it had no place in science. Positivists developed an elaborate view of the structure of scientific theory in which theoretical terms

were defined in terms of their ties with other theoretical terms and observables by scientific laws, without any reference to reality, involving no metaphysics. In brief, constructs were defined by relationships with other constructs, not by reference to reality. Cronbach and Meehl incorporated validation into the positivist framework and proposed that validity be determined by theory testing. Thus, construct validity theory could avoid realist claims about measured psychological constructs, and yet provide an explicit rigorous test of the validity of a construct via evidence for the network. Validity, then, is supported by the entire network: an ingenious idea!

However, this is problematic because for the most part there are no nomological networks in medical education (or psychology or education); there are no systems of scientific laws that explicitly link constructs and observables, and there is no theory of the construct to test, or at least nothing of the sort needed to establish construct validity. Originally, in laying out construct validity theory, Cronbach and Meehl emphasised that: 'To validate a claim that a test measures a construct, a nomological net surrounding the concept must exist.'[4] However, they also acknowledged the 'vagueness of present psychological laws' and said: 'Psychology works with crude, half-explicit formulations.'[4] The expectation was that eventually, with further research, theoretical concepts and their relationships would be clarified and an explicit theory (a nomological network) of psychology would emerge. Then the construct validity approach would become possible. Yet psychology doesn't seem to be any closer to this now than when it was first proposed.[9]

CONSTRUCT VALIDITY BASED ON INTERPRETATION AND ARGUMENT

To salvage the construct validity approach, seemingly less stringent criteria – interpretation and argument – have replaced nomological networks and rigorous theory testing for establishing validity. Messick, in opening his chapter on 'Validity' in the third edition of *Educational Measurement* (1989), wrote: '…what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that the interpretation entails.'[2] Similarly, in the fourth edition of *Educational Measurement* (2006), Kane concluded his chapter on 'Validation' by saying: 'Validation involves the evaluation of the proposed interpretations and uses of measurements. The interpretive argument provides an explicit statement of the inferences and assumptions inherent in the proposed interpretations and uses. The validity argument provides an evaluation of the coherence of the interpretive argument and of the plausibility of its inferences and assumptions.'[3] The current construct validity approach, then, seeks to establish a validity argument based on evidence for an interpretation of the target construct, but no longer within the framework of a rigorous nomological network that can 'fix the meaning of theoretical terms' in a way that can convincingly establish the validity of the construct.[7]

Kane recognised that: '…for validation to go forward, it is necessary that the proposed interpretations and uses be clearly stated.'[3] However, in practice, interpretation and argument seem to lack the 'glue' – the precision in prediction, testing and confirmation – needed to provide the confidence in the validity of the measurement of the postulated theoretical construct that was afforded by a nomological network. Researchers are left with vague, half-explicit formulations of the type that concerned Cronbach and Meehl 50 years ago. As Borsboom *et al.* wrote: 'The notion of a test score interpretation is too general.'[9] For example, inferences (interpretation and argument) commonly involve correlations between the construct and other variables, but, given that most variables are correlated with most other variables to some degree (especially with large enough samples),[16] correlations without an explicit theory are not informative about validity. Convergent and discriminant validity and multitrait–multimethod matrices[17] are commonly recommended for use with the current construct validity approach, but they require even more explicit theory to establish validity. At best, these validity arguments are weak, showing that one correlation is higher than another.[7]

The current construct validity approach seems to have come to focus more on reporting various 'sources of validity evidence'[18–20] (the current version of 'validity types', as recommended in the 'Standards for Educational and Psychological Testing'[18]), as if to compensate for the lack of nomological networks and the subsequent weakening of the theory testing part of construct validity by placing greater emphasis on 'evidence'. This seems to have diverted attention from the rationale and interpretation of the validity argument. Researchers attempting to validate a test then appear to list available evidence that fits into one of the 'sources' categories, but without showing how this supports the validity of the test (and at times it appears that it does not). That

is, all sorts of circumstantial evidence are cited for the interpretation/validity argument (such as improvement in scores with training, gender differences, internal consistency, number and names of factors or dimensions, and correlations with other variables). However, that evidence (females have higher or lower scores than males, or a three-factor structure versus a four-factor one, or clerks perform better than second-year students, etc.) does not establish directly with confidence that the instrument actually measures what it purports to measure (such as critical thinking, clinical reasoning, empathy, burnout, professionalism) and that the test is valid.

AND BACK?

Construct validity is an ingenious idea, but it has not lived up to expectations, primarily because explicit theory in psychology and education (and medical education) that would allow for the rigorous testing or validation of a measured construct is lacking. In the last decade, in a stimulating series of papers from the University of Amsterdam, Borsboom *et al.*[9] have considered the status of theoretical terms in psychology, in particular the construct validity approach, and concluded that this is 'the end of construct validity'. In 2009, they wrote: 'Psychology simply had no nomological networks of the sort positivism required in 1955, neither vague nor clear ones, just as it has none today. For this reason, the idea of construct validity was born dead … [it] never saw any research action.'[9]

In response to concerns about construct validity, these authors propose a *realist* approach to measurement (after the positivist ban), in which measurement is defined in terms of a *causal* relationship between variation in the attribute itself and variation in the measurement outcome or test score.[7–9] This 'realism and causal analysis' view sees 'the act of measurement as a product of a causal relationship between an instrument (broadly interpreted) and a magnitude': 'The magnitudes or quantities (properties, processes, states, events, etc.) exist independently of attempts to measure them.'[21] This thinking is not aimed at establishing new methods for the validation of a measurement instrument, but, rather, is more concerned with the definition of measurement in terms of *what can be measured* and *what counts as measurement.*

For all practical purposes, this somewhat abstract philosophical argument can be understood by thinking in terms of the distinction between the measurement of constructs versus the measurement of what might be called 'attributes'. Constructs, as discussed here, are abstract theoretical terms which are given their meaning by a nomological network or some approximation thereof (interpretation and argument) and exist only as ideas tied together with other ideas – hence the centrality of correlations in construct validation. Attributes, on the other hand are thought to exist apart from theory, and are measured by instruments for which outcomes are causally determined by the attribute. Attributes then are considered to be more than just theoretical ideas; rather, they are thought to exist independently of their measurement and serve to cause the measurement outcome.[21] For example, height, weight, blood pressure and scholastic performance can be implicitly or explicitly assumed to be attributes that are out there apart from measurement, and variations in these attributes cause variations in their measurements with a metre stick, pan balance, pressure cuff and grade point average, respectively. However, abstract theoretical constructs (like critical thinking, clinical reasoning, burnout, empathy, professionalism, systems-based practice, etc.) cannot convincingly be assumed to be out there apart from theory, and it is not clear that variation in their respective measurement instruments is caused by variation in the attributes.

Similar concerns are raised by Lurie *et al.*[22] about the assessment of competency-based educational objectives such as the core competencies proposed by the Accreditation Council for Graduate Medical Education.[23] Their concern is that educational competencies are 'political constructs' that are 'shaped by negotiations among stakeholders' and 'do not seem to have any demonstrated empirical basis'.[22] Philosopher John Searle makes a similar distinction in his writings on the construction of social reality, in which he distinguishes between 'brute facts' and 'social or institutional facts'.[24] The former refers to facts (attributes) that are thought to really exist out there, whereas the latter are acknowledged to be simply ideas or concepts that are limited to human thinking. Both are human social constructions, but the former has 'realist commitments' and the latter refers only to theory based on more theory.

This attribute-based view of measurement presented by Borsboom *et al.*[6–9] attempts to describe the essence of scientific measurement, not just a new validity type or a new theory of validity. Firstly, in these authors' words: 'If something does not exist, then one cannot measure it.'[7] Construct validity theorists proposed a positivist-based system to define and give existence to

an abstract theoretical construct by making reference to a network of other abstract theoretical constructs, but this, as discussed, has not been successful; otherwise, it is not clear in what sense constructs like these might exist. Secondly, Borsboom *et al.* write: 'Somewhere in the chain of events that occurs between item administration and item response, the measured attribute must play a causal role in determining what value the measurement outcomes will take.'[7] In brief, they are saying that measurement should be limited to 'attributes'.

Nevertheless, Borsboom *et al.*[6–9] use the term 'validity' to refer to measurements that meet these criteria: that is, if an attribute is thought to exist (independently of measurement) and causes the measurement outcomes, the instrument is said to be valid; otherwise, it does not measure the attribute and is not valid. This gives a different twist to the usual meaning of validity, which can be confusing. In addition, it makes validity into an all-or-nothing issue: either the instrument measures the attribute (and is valid) or it does not (and is not valid). Be that as it may, numerous extraneous factors may affect the measurement process in ways that add to the variability of the outcome measures. The sources of the added variability can be assessed with generalisability theory and analysis (i.e. this seems consistent with recent thinking about generalisability theory[25]). Consequently, an instrument may be valid, but its measurements not reliable. That is, an instrument may measure an attribute, but other factors in the measurement process may add irrelevant variance that affects the reliability of the measurements.

CONCLUSIONS

Construct validity has not proven to be a way to validate psychological constructs that have no clear referent in reality because explicit theory in medical education (and in psychology and education) that can provide a rigorous basis for validation is lacking. Interpretation and argument are not viable substitutes: simply listing any available evidence that fits in the various 'sources of validity evidence' categories recommended in the 'Standards'[17–19] does not show that the instrument measures what it purports to measure. It does not resolve the lack of nomological networks. Instead, it seems to weaken the concept of validity and to undermine the credibility of validity claims. The more general implication for medical education is that test developers and users should reconsider the value of using abstract theoretical constructs that have no referent apart from theory and that have no demonstrated empirical basis.[22,23] The use of the construct validity approach should be seriously reconsidered for research in medical education.

Assessment research and practice in medical education might be better served by more modest concrete indicators (attributes) that are often readily available and standard across training and practice, especially given the extensive record keeping in medical education. The primary purpose of research in medical education does not seem to be to establish an abstract psychological-type theory that consists of abstract psychological-type constructs, but, rather, is more practical and should be aimed at determining relationships among basic variables or measurements that can be used to better understand teaching and learning in medicine.[22] This suggests that research should concentrate on areas of study more than on constructs to validate. For example, research in the area of professionalism provides valuable results about relationships between information in medical school records and state board disciplinary action, which is very important research but does not require the postulation of a construct of professionalism in order to do so.[26] A focus on such basic measures (and areas of research) would avoid the problems associated with (and perhaps the impossibility of) establishing the validity of abstract theoretical constructs. It may also reveal that the development of long sought-after theory in medical education is better served by building theory from the bottom up rather than from the top down and by then combining the results of multiple studies and explaining them with higher-order constructs developed for that purpose.[22]

Contributors: all authors contributed equally to the conception and design of this paper, and to the drafting and critical revision of the manuscript. All authors approved the final manuscript for submission.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Brennan RL. Perspectives on the evolution and future of educational measurement. In: Brennan RL, ed. *Educational Measurement*, 4th edn. Westport, CT: American Council on Education/Praeger 2006;1–60.
2 Messick S. Validity. In: Linn RL, ed. *Educational Measurement*, 3rd edn. New York, NY: American Council on Education/Macmillan 1989;13–103.

3 Kane M. Validation. In: Brennan RL, ed. *Educational Measurement*, 4th edn. Westport, CT: American Council on Education/Praeger 2006;17–64.
4 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.
5 Norman G, Eva K. Quantitative research methods in medical education. In: Swanwick T, ed. *Understanding Medical Education: Evidence, Theory, and Practice*. Oxford: Wiley-Blackwell 2010;301–22.
6 Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. *Psychol Rev* 2003;**110**:203–19.
7 Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. *Psychol Rev* 2004;**111**:1061–71.
8 Borsboom D. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. New York, NY: Cambridge University Press 2005.
9 Borsboom D, Cramer AOJ, Kievit RA, Scholten AZ, Franic J. The end of construct validity. In: Lissitz RW, ed. *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte, NC: Information Age Publishing 2009;135–70.
10 Kelly TL. *Interpretation of Educational Measurements*. New York, NY: MacMillan 1927;14.
11 Anastasi A. *Psychological Testing*. New York, NY: MacMillan 1954;29.
12 Kane MT. Current concerns in validity theory. *J Educ Meas* 2001;**38**:319–42.
13 American Psychological Association. Technical recommendations for psychological test and diagnostic techniques. *Psychol Bull Suppl* 1954;**51**:1–38.
14 Landy FJ. Stamp collecting versus science: validation as hypothesis testing. *Am Psychol* 1986;**41**:1183–92.
15 Suppe F. *The Structure of Scientific Theories*. Urbana, IL: University of Illinois Press 1977.
16 Hayes WL. *Statistics for Psychologists*. New York, NY: Holt, Rinehart & Winston 1963;326.
17 Campbell DT, Fiske DW. Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychol Bull* 1959;**56**:81–105.
18 American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association 1999.
19 Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003;**37**:830–7.
20 Varkey P, Natt N, Lesnick T, Downing S, Yudkowsky R. Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: a preliminary investigation. *Acad Med* 2008;**83**:775–80.
21 Trout JD. Measurement. In: Newton-Smith WH, ed. *A Companion to the Philosophy of Science*. Malden, MA: Blackwell Publishers 2000;265–76.
22 Lurie SJ, Mooney CJ, Lyness JM. Pitfalls in assessment of competency-based educational objectives. *Acad Med* 2011;**86**:412–4.
23 Accreditation Council for Graduate Medical Education. ACGME General Competencies and Outcomes Assessment for Designated Institutional Officials. http://www.acgme.org/acWebsite/irc/irc_competencies.asp.
24 Searle J. *The Construction of Social Reality*. New York, NY: Free Press 1995.
25 Cardinet J, Johnson S, Pini G. *Applying Generalizability Theory using EduG*. New York, NY: Taylor & Francis Group 2010.
26 Papadakis MA, Hodgson CS, Teherani A, Kohatsu ND. Unprofessional behaviour in medical school is associated with subsequent disciplinary action. *Acad Med* 2004;**79**:244–9.

Received 5 July 2011; editorial comments to authors 22 September 2011, 3 November 2011; accepted for publication 17 November 2011