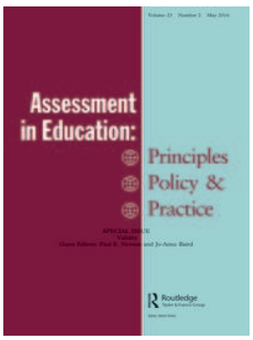

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

Disagreement over the best way to use the word 'validity' and options for reaching consensus

Paul E. Newton & Stuart D. Shaw

To cite this article: Paul E. Newton & Stuart D. Shaw (2016) Disagreement over the best way to use the word 'validity' and options for reaching consensus, Assessment in Education: Principles, Policy & Practice, 23:2, 178-197, DOI: 10.1080/0969594X.2015.1037241

To link to this article: http://dx.doi.org/10.1080/0969594X.2015.1037241

Assessment in Education: Principles, Policy & Practice, 2016
Vol. 23, No. 2, 178–197, http://dx.doi.org/10.1080/0969594X.2015.1037241

Disagreement over the best way to use the word 'validity' and options for reaching consensus

Paul E. Newton[a]* and Stuart D. Shaw[b]

[a]*Office of Qualifications and Examinations Regulation, Coventry, UK;* [b]*Cambridge International Examinations, Cambridge, UK*

(*Received 27 October 2014; accepted 31 March 2015*)

The ability to convey shared meaning with minimal ambiguity is highly desirable for technical terms within disciplines and professions. Unfortunately, there is no widespread professional consensus over the meaning of the word 'validity' as it pertains to educational and psychological testing. After illustrating the nature and extent of disagreement, we consider three options for reaching consensus: to eliminate its ambiguity by agreeing a precise technical definition; to embrace its ambiguity by agreeing a catchall lay usage; and to retire 'validity' from the testing lexicon.

Keywords: validity; validation; quality; value

Preface

Validity is widely acknowledged to be the most fundamental consideration in developing and evaluating tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014, p. 11). Yet, if one were to gather a sample of psychometricians in a bar, 'a mêlée would ensue if they were asked to define what validity *is*' (Forte Fast & Hebbler, 2004, p. i). How could that be?

According to two prominent validity scholars, validity is so simple that even 'a bright 8-year-old could grasp the general idea' (Borsboom & Mellenbergh, 2007, p. 93). Yet, according to another, her undergraduate class on validity is 'the most challenging class of the semester' and the hardest 'for students to understand' (Gorin, 2007, p. 456). Again, how could that be?

The multitude of conflicting perspectives, which we will discuss in more detail below, clearly illustrates that validity has become 'a conceptual animal at war with itself' (Baker, 2013, p. 22). But how has this conflict arisen? And what are the prospects for achieving peace and reconciliation? These are the kinds of question that prompted us to write this paper.

Consensus is a good thing

During an early meeting of the National Association of Directors of Educational Research (26 February 1920, Cleveland, Ohio, USA), a permanent committee was

*Corresponding author. Email: newton.paul@virginmedia.com

established to address the issue of standardisation. The following year, its chairman presented findings from a questionnaire that had been sent to association members, reporting 'a practically unanimous sentiment in favor of the publication of an official list of terms, procedures, etc.' (Buckingham et al., 1921, p. 80; emphasis added). The Committee prepared a provisional set of official definitions, urging members to conform to them whilst preparing material for publication. It defined the 'problem of validity' as the 'determination of what a test measures' (Buckingham et al., 1921, p. 80).

The fact that it is a good thing for members of a professional community to attach the same meaning to the technical terms through which they communicate on a daily basis ought to go without saying. It is obviously a good thing. Words that mean quite different things to different members of a community erect a barrier to effective communication. Unfortunately, the word 'validity' presents exactly this problem to members of the educational and psychological testing community. As the most hallowed term in our lexicon, this seems almost ironic.

The lack of consensus is certainly not for want of trying. Over the past century, many individual scholars have attempted to persuade others of the benefits of thinking and talking about validity in one particular way or another (see Newton & Shaw, 2013, 2014). More importantly, committees and joint committees of professional organisations have collaborated determinedly to help establish consensus over the meaning of this elusive word. Successive editions of the North American *Standards* illustrate this point most clearly.[1] Yet, no widespread professional consensus has yet been achieved.

The purpose of this paper is to persuade the reader that it is both interesting and important to consider why consensus has not been achieved, and to consider whether anything realistically can be done about it. We therefore start from the premise that a greater degree of consensus over how best to use the word 'validity' is a worthwhile ambition. In particular, this paper is written for anyone who might otherwise have assumed that:

(1) consensus over the best way to use the word 'validity' has already been established; or that
(2) we seem to be converging upon a widespread professional consensus and are closer to it now than ever before; or that
(3) despite superficial terminological dissimilarities, all of the existing definitions say more-or-less the same thing.

We will demonstrate that none of these three statements is true and that the situation is better described as a standoff between scholars (and their followers) who advocate radically different usages. Admittedly, toward the end of the twentieth century, it had begun to seem as though consensus might finally have been achieved (e.g. Dunnette, 1992, p. 160; Shepard, 1993, p. 405; Moss, 1995, p. 6). Yet, developments during the first decade of the twenty-first century revealed that this impression was far from true (e.g. Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2004; Cizek, 2012; Lissitz & Samuelsen, 2007; Lissitz, 2009; Newton, 2012a, 2012b; Newton & Shaw, 2013). We begin by illustrating the nature and extent of disagreement.

Lack of consensus abounds

Connotations of the word 'validity' have been debated ever since it entered the official lexicon of educational and psychological testing during the 1920s. Debate has intensified recently and shows no sign of abating. There are multiple dimensions of disagreement, but the two highest-profile controversies concern: (i) what validity should encompass; and (ii) what validity should apply to. For reasons of space, we will construct our argument on the basis of the first controversy alone.[2]

The first of these two controversies, which concerns what validity should encompass, became especially prominent in the wake of seminal papers by Samuel Messick (e.g. 1980, 1989a). Many read his work to be arguing for a reform of validity theory, which would extend its scope considerably beyond the traditional scientific concern for good measurement, to include a new ethical concern for good consequences. We label those reformers who marshal under the 'consequential validity' banner the new liberals. We contrast them with traditionalists, whose view of validity embraces both the science of measurement interpretation and the technology of predictive use. This perspective reflects a received wisdom that has been passed down through successive editions of the *Standards*; particularly the first four editions.[3] Interestingly, as critics of the new liberal position made their views known, many of them insisted that the scope of validity should actually be far narrower than even traditionalists claim. These critics hark back to the very first definitions of validity, which were framed *purely* in terms of measurement science. We therefore describe them as the new conservatives. The most extreme example of this critical perspective is important enough to be considered separately. It is best described as ultra-conservative, because it seeks to strip-back validity to its barest measurement foundation. We summarise these contrasting perspectives below.

Traditionalist

Although validity was originally defined in terms of what a test measures and was typically judged by correlating test scores against a concurrent criterion, it soon became apparent that testing was not simply about measurement and that the so-called 'validity coefficient' was not always ideal for judging validity (e.g. Pressey, 1920). Indeed, for many early testing professionals, the ultimate proof of a test used for predicting future outcomes was simply its predictive power, regardless of whether it could be said to measure anything at all.

Traditionalists, in keeping with successive editions of the *Standards*, presume that validity should somehow encompass both measurement (which focuses attention on test score interpretation) *and* prediction (which focuses attention on test score use). This scope was very clear in the first edition of the *Standards* which indicated that:

> Validity information indicates to the test user the degree to which the test is capable of achieving certain aims. (APA, AERA, & NCMUE, 1954, p. 13)

The four aims identified by the *Standards* included both measurement-focused ones (e.g. those requiring evidence of content validity) and prediction-focused ones (e.g. those requiring evidence of predictive validity). Subsequent theoretical work, including seminal chapters by Cronbach (1971) and Messick (1989a), emphasised that it was unhelpful to think of test score interpretation and test score use as distinct aims.

Instead, the broad constructs through which test scores are interpreted provide a warrant for using test scores across a variety of situations. Moreover, if the use of a test score to make a particular decision necessitates a particular kind of interpretation, then it is this interpretation that requires validation, regardless of what the test might originally have been designed to measure (for further elaboration, see Newton, 2012a; Shepard, 1990, 1997).

Testing, from the traditionalist perspective, is not a pure science. Validity is inherently pragmatic. Tests are not created in a vacuum; they are created for a purpose. Test scores are not generated simply to be interpreted; they are generated to be used. Indeed, from the traditionalist perspective, one might even go so far as to claim that it is principally the *use* of test scores that needs to be validated (Shepard, 1993; Sireci, 2007).

The growing significance of values

It is tempting to think that validity, from the traditionalist perspective, is essentially technical (if not ultimately scientific) in outlook. However, it was recognised long ago that even tests with low predictive power can still be very useful under certain circumstances (Taylor & Russell, 1939). This idea of usefulness extends beyond the technical outlook, to highlight economic, social and ethical concerns, framed squarely in terms of costs and benefits.

Usefulness has often been described in the testing literature using the word 'utility' and theorised independently of validity (Cronbach & Gleser, 1957, 1965). However, this conceptual bracketing would be problematic for anyone who insisted that validity is as much a property of decision-making (test use) as measurement (test interpretation). This might suggest that the traditionalist perspective on validity ought somehow to be able to encompass the kind of value judgements that are required to establish test usefulness.

As attention turned from earlier work on financial utility, to later work which focused on moral fairness, the need to embrace social and ethical considerations became even more apparent (e.g. Cronbach, 1976; Cronbach, Yalow, & Schaeffer, 1980).

Liberal

A new liberal perspective has remained at the centre of the past few decades of debate over the significance of consequences to validity and validation. It fully embraces the idea that it is insufficient, if not irresponsible, to evaluate tests from a purely scientific or technical perspective. Rather than treating validity as the overarching evaluative concept for addressing measurement quality, or even decision-making value, it treats validity as the overarching evaluative concept for addressing the overall defensibility, or acceptability, of the policy which sanctions the use of test scores for a particular purpose.

The new liberal perspective can be viewed as an extension of the traditionalist perspective. Whereas traditionalists very soon extended the classic definition of validity to encompass both measurement (test score interpretation) and prediction (test score use), liberals extended traditionalist definitions to encompass both intended and unintended consequences arising from test score use. Whereas questions of social value clearly arise when making decisions on the basis of test scores,

they become the central focus when considering the full range of consequences associated with test-based decision-making.

Interestingly, although this perspective has received a lot of attention – the idea of 'consequential validity' came in for a lot of criticism during the 1990s, e.g. Maguire, Hattie, and Haig (1994), Norris (1995), Lees-Haley (1996), Sackett (1998), Tenopyr (1998) – it is actually quite tricky to identify ardent liberals unambiguously. Many scholars nowadays seem to have distinct liberal leanings, but few seem to nail their colours to the mast definitively. Early papers on validity by Messick seem to promote a liberal perspective (e.g. Messick, 1980). Yet, his later papers seem far more traditionalist in outlook (Newton & Shaw, 2014). Later papers by Cronbach seem to embody the new liberal perspective, especially Cronbach (1988). It can also be detected in the work of Moss (1992, 1995, 1998) and Linn (1993, 1997).

Conservative

A new conservative perspective has become increasingly prominent in recent years. It represents a return to the classic definition of validity, framed in terms of 'the degree to which a test or examination measures what it purports to measure' (Ruch, 1924, p. 13; see also Buckingham, 1921, p. 274). Proponents of this perspective argue forcefully that validity is not a pragmatic concept but a scientific one. The term can therefore only properly be applied to measurement and to measuring procedures (e.g. Cizek, 2012; Scriven, 2002). It is a category error, from this perspective, to refer to valid predictions, valid uses of results or valid decision-making procedures, let alone to a valid testing policy. As Cizek recently put it, test interpretation and test use are 'incompatible concerns' (2012, p. 31):

> The first endeavour is one that gathers and evaluates support for test score inferences; that is, *validation*. The second endeavour is one that gathers and evaluates support for test use; that is, *justification*. (Cizek, 2012, p. 41)

Many proponents of the new Conservative perspective see validity as an umbrella concept, albeit a fairly narrow one, focused specifically upon measurement quality. Included within this concept are many distinguishable subsidiary concepts, like reliability/precision, dimensionality and bias.

Ultra-conservative

The second edition of an early glossary of testing terms illustrated a conceptual debate that has continued to the present day: should reliability be considered a subsidiary concept within validity, a parallel concept alongside it or simply two aspects of the same concept (Odell, 1928; see also Campbell & Fiske, 1959; Kane, 2006; Marcoulides, 2004; Messick, 1989a; Thurstone, 1931)?

> Thus a test cannot be valid unless it is objective and reliable, but can be perfectly objective and reliable without being valid. […] It has also been suggested that the term valid should be used in a more restricted sense than that just explained. In this sense it would exclude the factor of reliability. (Odell, 1928, p. 65)

Over the past decade or so, Borsboom and colleagues have recommended a return to the classic definition of validity (Borsboom, 2005; Borsboom et al., 2004, 2009).

However, they have argued for a more extreme retreat than other conservatives, defining validity as nothing more than the most fundamental requirement for the truth of any measurement claim: that the measurand (that which is supposedly being measured) is causally responsible for variation in outcomes from the measuring procedure. If it is causally responsible, then it is true that the procedure is valid for measuring what it is supposed to measure. If not, then validity cannot legitimately be claimed.

Importantly, the concept of validity, from this perspective, is very narrow, encompassing far less than any other perspective. Validity is defined independently of other evaluative concepts, like reliability/precision, dimensionality and bias. Indeed, from this perspective, a measuring procedure can be both valid *and* biased, just as it can be both valid *and* unreliable. Definitional independence is claimed to be a particular virtue. The conservative and ultra-conservative positions have been contrasted thus:

> Thus, it appears that one can view validity either as an integrative function of psychometric properties, or as a separate property that is orthogonal to psychometric functioning. This is an important issue, because one's viewpoint here determines what one takes as validity evidence. On the integrative viewpoint, for instance, high measurement precision would count as one piece of evidence for validity by itself (as validity is an overarching property which has reliability as one [of] its constituents). In the orthogonal viewpoint, it would not, as high reliability is neither necessary nor sufficient for validity. Reliability is not necessary, because one may have an instrument that does in fact measure the intended attribute, but does do with low measurement precision. (Markus & Borsboom, 2013, p. 64)

Non-convergence and incompatibility

Based upon the evidence so far presented, we think that it is reasonable to conclude that consensus over the best way to use the word 'validity' has not yet been established. To be fair, there are no large-scale surveys of professional opinion to warrant generalising this conclusion across the entire international community of testing professionals. However, it would be surprising if the lack of agreement amongst scholars was not somehow reflected more broadly. Divergent perspectives are clearly apparent across a wide range of published resources. Textbooks provide an excellent illustration of this, helping to explain why validity is so very complicated to teach and to learn about:

> A test is said to be valid if it measures what it purports to measure. (Kline, 1998, p. 34)

> a test score is *valid* to the extent that it measures the attribute of the respondents that the test is employed to measure, in the population(s) for which the test is used. (McDonald, 1999, p. 197)

> At its essence, validity means that the information yielded by a test is appropriate, meaningful, and useful for decision making – the purpose of mental measurement. (Osterlind, 2010, p. 89)

> **Validity** is defined as the extent to which measurements are useful in making decisions and providing explanations relevant to a given purpose. (Sax, 1997, p. 304)

> Validity is the adequacy and appropriateness of the interpretations and uses of assessment results. (Miller, Linn, & Gronlund, 2009, p. 70)

> validity refers to the meaningfulness and defensibility of the actions or decisions based on test scores, test-based information or assessment reports. (Chatterji, 2013, p. 275)

The definitions are not just superficially different; they clearly transform from conservative at the top, through traditionalist in the middle, to liberal at the bottom. McDonald, for instance, explicitly excluded prediction from his definition, relabeling predictive 'validity' predictive 'utility' (McDonald, 1999, p. 199). Yet, Sax made utility, or usefulness, the foundation for his definition. The definitions provided by Miller et al. and Chatterji appear to go even further, embracing two heavily value-laden concepts, appropriateness and defensibility. Primary sources on validity read even more adamantly concerning the incompatibility of competing perspectives:

> Their mistake, I believe, is in trying to tie social consequences into a validity framework. Such a wedding of related but distinctive concepts will not be symbiotic, it will be septic. (Popham, 1997, p. 13)

> Although Messick wants to move to what he calls unified validity, he takes this to include both of what are, I suggest, properly called validity and utility. (Scriven, 2002, p. 259)

> I conclude with a word on what can be learned regarding *consequential validity* from these deliberations. […] The chapters in this book lead me to conclude that *validity* and *test use* considerations are inseparable from *consequences*. (Chatterji, 2013, p. 306)

> Put simply, consequential validity doesn't exist. (Cizek, 2010, p. 4)

Considering evidence like this, the idea that, as a testing community, we are somehow converging upon a widespread professional consensus over the best way to use the word 'validity' seems far-fetched. Indeed, just as there are signs of increasing sympathy for a more conservative perspective, so there are signs of increasing sympathy for a more liberal one (e.g. Brennan, 2013, p. 80). As these opposing perspectives acquire an increasing number of converts, so it appears that the testing community diverges further from consensus than ever before.

A matter of logic or a matter of consequence?

Disagreements over how best to use the word 'validity' often seem to be constructed so as to give the impression that they can be resolved on purely logical grounds, i.e. that one view is straightforwardly right and others straightforwardly wrong. Occasionally, certain elements of certain debates do seem to be amenable to this kind of resolution. If, for instance, we start from the classic definition of validity – the technical quality of a measuring procedure – then logic alone is sufficient to explain how evidence from the consequences of testing has the potential to support or challenge any prior claim to validity (Messick, 1989a; Shepard, 1997).

Unfortunately, many disagreements over the use of the word 'validity' are far more than straightforward matters of logic, and can only be arbitrated on consequential grounds, concerning the consequences of using the word 'validity' in one way rather than another. For example, concerning the controversy over what 'validity' should apply to, a consequential argument for not referring to 'the validity of the test' is the high risk of confusing test users into thinking that the test has been validated unconditionally, and the consequent risk that test scores may be used for pur-

poses for which they are not fit (Frisbie, 2005; Newton, 2012b; Newton & Shaw, 2013).

Interestingly, similar argument structures have been employed to defend competing perspectives. Thus, an important consequential argument for a more liberal perspective focused on risks from *excluding* consequences from the concept of validity: in particular, the risk that a positive evaluation of test score interpretation, i.e. good measurement, might be misinterpreted as a prima facie justification of test use (Kane, 2013a, p. 62). Yet, an equally important consequential argument for a more conservative perspective focused on risks from *including* consequences within the concept of validity: in particular, the risk that a negative evaluation of test score use, i.e. bad decision-making, might be misinterpreted as a prima facie negation of test score interpretation (Mehrens, 1997, p. 17).

Whether it is best to use 'validity' in a more liberal sense, a more conservative sense or in any other sense, is not straightforward a matter of logic. It is a matter of consequence, concerning the costs and benefits of adopting one perspective rather than any other. Significantly, there are very compelling cost-benefit arguments on both sides. We will return to this point shortly, after considering a number of possible objections to our starting premise.

Forget consensus

We believe that a widespread professional consensus over how best to use the word 'validity' would be a good thing. Ideally, this would involve a precise, technical definition. Our premise is very simple: common usage facilitates effective communication; divergent usages hinder it. However, we have encountered various responses to our premise, which we briefly address below.

Objection 1: family resemblance

The first objection recommends that there should be no ambition for consensus over a precise, technical definition because the word 'validity' does not afford this level of conceptual clarity. It is a family resemblance concept for which there could be no precise definition. The best that we can hope for is a loose, implicit consensus over the proper application of the term.

Contrary to terms that can be defined precisely, by specifying the feature(s) which they share in common, family resemblance terms function by virtue of multiple features, none of which are shared by all instances, but which are shared across instances in a 'criss-crossing' manner (Forster, 2010). What enables a family resemblance concept to overcome its resistance to precise *definition* is that there exists sufficient intersubjective agreement over its *application*.

Therein, of course, lay the rub. The fact that there is so much explicit disagreement over how best to use the word argues strongly against its status as a family resemblance concept (Newton & Shaw, 2013). We would be happy acknowledge and embrace a family resemblance consensus over the meaning of validity – if one could be said to exist. But, instead, we are faced with a choice between explicitly competing perspectives, which is an entirely different ballgame.

Objection 2: state of flux

A second objection insists that there should be no ambition for widespread consensus – and certainly no attempt to regulate the language of testing – because the science of testing will always be in a state of flux, and fundamental disagreement is the engine of progress in science. To regulate use of the word 'validity' would be to stifle progress in understanding testing. This is to argue that insisting upon a particular definition of validity is like the seventeenth-century declaration by the Catholic Church that the earth is the centre of the universe, and that any definition to the contrary is heresy.

We note, however, that even scientific communities are far from averse to upholding particular views of the world as 'the truth' if only provisionally or for pragmatic reasons. The historical definition and redefinition of the metre provides a good example of this, including the *Metre of the Archives* in 1799, the cross-national *Treaty of the Metre* in 1875 and so on (Penzes, n.d.).

There would be more force to this objection if debate over how best to use the word 'validity' reflected fundamental disagreements over the core evaluative concepts of testing. However, as we will shortly explain, that this is not the case. In a very important sense, the word 'validity' is just a word; a label for a concept. Labels are matters of convention rather than truth; which, again, supports the presumption that consensus is quite fundamental.

Objection 3: essentially contested

A third objection recommends that there should be no ambition for consensus, because contest is the hallmark of a concept like validity – an essentially contested concept (Gallie, 1956). Essentially contested concepts characteristically involve the appraisal of complex domains in terms of multiple criteria. Although there may be general agreement over the criteria at stake – supporting the claim that we are dealing with a single concept – different groups will weight those criteria differently and, more importantly, each group will promote their own approach to appraisal as the *true* embodiment of the concept. Dispute over what 'validity' should encompass might therefore be understood as the inevitable manifestation of a concept like this; with certain groups elevating score interpretation above score use and others doing the reverse. If validity is a concept like this, then perhaps there can be no hope of consensus, period.

We would argue that validity is not an essentially contested concept in the sense described by Gallie (1956). First, the debate over what 'validity' should encompass has been characterised by fundamental disagreement over the relevance of evaluative criteria – some rejecting ethical evaluation and some embracing it – which suggests that protagonists may not have been disputing a single concept, but confusing separate concepts. Second, the debate cannot actually be explained in terms of fundamental differences between groups in terms of their underlying values. No one would seriously question the importance of ethical evaluation to testing; just as no one would seriously question the importance of scientific evaluation. Third, a defining feature of essentially contested concepts is that their ambiguity cannot be resolved by stipulation, because no stipulator would ever be universally recognised. Yet, the remarkable tenacity of organisations like AERA, APA and NCME – in pro-

ducing successive editions of the *Standards* – suggests at least considerable momentum in this direction.

Prospects for reaching consensus

We have argued that a widespread consensus over a (fairly) precise meaning for the word 'validity' would be important to maximise its utility as a technical term. We now consider prospects for reaching this kind of consensus; or, indeed, for reaching any kind of consensus. We will consider three alternatives.

Option A: eliminate ambiguity

Faced by mounting evidence of disagreement, some might insist that we must redouble efforts to agree upon a technical definition: if we try hard enough, then rational debate will eventually lead us to consensus over a (fairly) precise meaning. We believe that prospects for this option may be poor, as there is already a great deal of empirical evidence against it, from nearly a century of impassioned debate (Newton & Shaw, 2013).

Closer inspection of this history may help us to understand why consensus is likely to remain elusive. Consider the argument between conservative and liberal perspectives, which has been conducted primarily on consequential grounds. From the liberal perspective, excluding ethical evaluation from validity risks no one taking responsibility for it. This seems to be what Messick meant when he reflected upon a passage written by Cronbach, who had used three examples of negative consequences arising from essentially truthful measurements to argue that ethical considerations were fundamental to validation:

> The bottom line is that validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences (Messick, 1980). You [...] may prefer to exclude reflection on consequences from the meanings of the word *validation*, but you cannot deny the obligation. (Cronbach, 1988, p. 6)

> But we would prefer a somewhat stronger phrasing, because the meaning of validation should not be considered a preference. On what can the legitimacy of the obligation to appraise social consequences of test interpretation and use be based, we argue, if not on the only genuine imperative in testing, namely, validity? (Messick, 1989b, p. 11)

This exchange reveals how Messick recognised the extraordinary power of the word 'validity' (within the testing community) in signalling what constitutes good practice in testing. The exclusion of ethical implications from the word 'validity' might risk appearing to absolve testing evaluators of any responsibility for investigating adverse consequences.

From the conservative perspective, including ethical evaluation within validity risks making it needlessly complicated; or, alternatively, too gross to be useful. Popham (1997) emphasised that conventional conceptions of validity were complicated enough, and that any extension of meaning would make it extremely hard for educators to understand the word. Mehrens, defending a similar position, observed:

> Words are powerful. How we use them is important. […] If validity is everything, then validity is nothing. (Mehrens, 1997, p. 18)

In other words, to broaden the scope of 'validity' substantially, potentially to embrace the overall acceptability of a testing policy, would render the word impotent. Most fundamentally, it could no longer function as a useful tool with which to highlight the critical features of high-quality measuring procedures (see also Wiley, 1991).

In presenting these examples, our intention is to highlight the emotional significance of the word 'validity' to individuals, organisations and communities of scientists and professionals involved in testing. It is not trivial that validity is repeatedly publicly lauded as the:

- 'most important examination quality' (Association of Language Testers in Europe, 2001),
- 'key criterion driving assessment' (Cambridge Assessment, 2009), and
- 'most fundamental consideration in developing tests and evaluating tests' (AERA, APA, & NCME, 2014).

The word 'validity' is treated with adulation. It has become a watchword, a slogan and a rallying cry. It is not just *a* word; it is *the* word, *our* word. As such, it somehow needs to be capable of satisfying everybody in the field of testing, whatever their particular role, circumstance or value base. Paradoxically, this word may have become *too* important for us. Its signalling function – focusing attention on what really matters when it comes to good testing – means that it really must cover all bases, from scientific to ethical (see Messick). Yet, if it does cover all bases, then it becomes everything which risks it becoming nothing (see Mehrens). It seems too risky for validity to embrace everything, when it comes to good testing; but, equally, it seems too risky for validity not to. The strength of the arguments in support of opposing perspectives frustrates any rational approach to conflict resolution. Indeed, the great validity debate of the twentieth century (Crocker, 1997) may well have degenerated into the great validity stalemate of the twenty-first century. The more evenly fought the conflicts, from a rational point of view, the higher emotions are likely to rise. The word 'validity' has perhaps become a stumbling block in its own right, a hostage to fortune, a trigger of psychometric mêlée. Prospects for reaching consensus over a precise technical definition for validity seem to be very low indeed.

Option B: embrace ambiguity

Contrary to the aspiration for a precise technical definition, we note the imprecision with which the word 'validity' appears (to us) so frequently to be used in everyday conversation, even between testing specialists (ourselves included). Theorists may lose sleep over whether the word 'validity' can legitimately be applied to an item, a score, a test, a measuring procedure, a use or decision, a decision-making procedure, a consequence or impact, a policy, a claim, a conclusion, an interpretation, an argument, an inference, an explanation or a theory. Practitioners, however, do not.[4] There is an uncomfortable disjunction between the historical aspiration for precision and the omnipresent reality of imprecision.

Against this backdrop, the best solution might simply be to accept that the word 'validity' has *already* become everything, but in a more extreme sense than even many liberals would recommend. Conceivably, despite the warning from Mehrens,

this might not actually be such a bad thing. This is to propose retaining the word 'validity' as a pre-eminent term within educational and psychological testing, but with the barest of consensus meanings possible. It should have no precise technical definition; conveying, instead, nothing more than a positive evaluation in relation to any aspect of testing – a valid item, a valid instrument, a valid measuring procedure, a valid policy and so on. This is to recommend moulding it into a transcendent family resemblance concept. Its usefulness would lie in allowing specialists to communicate effectively with non-specialists, at a level of informal generality that does not require precision of meaning or that might even be hindered by a requirement for precision. It would be equally useful in allowing specialists to converse with specialists, when precision of meaning was not required. If precision were to be required, then more clearly delineated technical terms could be employed.

The catchall lay term 'health' functions very much like this. This word derived from the old English word 'hoelth' which had connotations of being sound or whole (Simmons, 1989). Nowadays, the term can legitimately be used to convey a positive evaluation in relation to any aspect of living: a healthy toe, healthy skin, a healthy body, a healthy lifestyle, a healthy mind and so on.[5]

Option B, then, recommends that we recognise the word 'validity' as the most important testing concept by virtue of its lay significance; allowing issues of technical quality and social value to be productively discussed with even the most assessment illiterate of consumers and stakeholders. Anticipating an obvious response – that this would trivialise the concept – we note that productive discussion with customers and stakeholders is far from a trivial matter; it is fundamental to improving public understanding and confidence in testing.

We think that Option B is, in principle, very attractive. However, we also recognise how emotionally attached to the word 'validity' members of the testing community are wont to be. We readily acknowledge how challenging it would be for our community to agree to retain the term whilst draining it of all but the most general of meanings. Counter-intuitively, it might in fact prove easier to reach consensus *not* to use the word 'validity' *at all*. This is our third option.

Option C: retire 'validity'

> Maybe the term *validity* has outlived its usefulness. I would be happy to retire the word to the scrapheap of overused terms dying of terminal ambiguity (but I have not yet done so). (Guion, 2011, p. 181)

Anyone who presumed that lack of consensus reflected fundamental disagreements over core evaluative concepts might well consider the proposal to retire it akin to an ostrich sticking its head in the sand. Those debates would surely continue and we would simply end up disagreeing over how best to use different words. In fact, though, this is not at all how we see the lack of consensus.

We would certainly not deny the existence of fundamental disagreements over substantive testing concepts. There are fundamental debates concerning the meaning of 'measurement' when applied to educational and psychological testing; in particular, whether educational and psychological attributes can be presumed to possess the structure implied by a strong measurement interpretation, or whether the term can legitimately be employed with only a weak interpretation (Finkelstein, 2003, 2009; Michell, 1999, 2009). There are even more fundamental debates concerning the

ontology of the attributes that we presume to be measuring (Hood, 2009; Maraun, 1998; Maraun, Slaney, & Gabriel, 2009; Maul, 2013). Yet, debates concerning how best to use the word 'validity' tend to be considerably less substantial than these.

One recurrent theme in these debates is the question of responsibility. For instance, given that test developers and publishers have traditionally owned the concept of validity, would extending its scope to consideration of consequences render them responsible for more than they can reasonably be held responsible (Green, 1998; Reckase, 1998; Urbina, 2004)? Although the question of responsibility is undeniably important, it is not actually the kind of question that could be resolved by consensus over the meaning of a word. Even the most liberal of definitions of 'validity' says nothing about who ought to be responsible for which aspects of validation. Indeed, those who adopt a more liberal perspective often emphasise the importance of sharing validation responsibilities widely (Haertel, 2013; Linn, 1998).

More generally, despite substantial disagreement concerning whether or not the word 'validity' ought to have ethical connotations, no one disagrees that comprehensive evaluation of testing requires consideration of ethical issues. In the same way, there is considerable consensus over the core concepts with which to understand comprehensive evaluation. Few, we presume, would deny the importance of evaluating the three core objectives of testing: measurement objectives, albeit widely defined measurement; decision-making objectives; and secondary policy objectives concerning broader impacts. Few, we presume, would deny the importance of evaluating each of these objectives from both scientific and ethical perspectives; that is, in terms of both technical quality and social value. These, we believe, encapsulate the core substantive evaluative concepts of educational and psychological testing (Newton & Shaw, 2014, chapter 6). Disagreement over the word 'validity' has not focused on the inadequacy or inappropriateness of these concepts, per se; although there has been much debate over which of them ought, and ought not, to be associated with the label. *In other words, disagreement has focused primarily upon how best to apply the label, not upon how best to apprehend the underlying concepts*. If we were to retire the word 'validity' our substantive concepts would undoubtedly survive intact.[6] Indeed, given the omnipresent reality of imprecise and ambiguous usage, and the fact that even testing specialists use the word in quite different ways, it is hard to see how anything *conceptually* fundamental could be lost if the word were to be retired. The most unfortunate consequences of the current debate are that its adversarial nature: (a) encourages the taking of sides, when there are important lessons to be learned from all perspectives; and (b) makes it appear that the protagonists are somehow fundamentally opposed in terms of their outlook on the nature and scope of evaluation within testing, which is very far from the truth.

Option C resonates with the recommendation, from Mameli and Bateson, to retire the term 'innateness' from scientific discourse because it has become cluttered through association with all sorts of different properties that cannot be assumed to cluster together (Mameli & Bateson, 2006, 2011). They dismissed the alternative of stipulating a definition, i.e. associating the term with one particular property – akin to our Option A – because this would arbitrarily rule-out many of the inferences and classifications that scientists routinely associate with 'innateness' and because of the confusion that would inevitably ensue.

Finally, we note with interest that Bachman and Palmer (2010), a 510-page textbook entitled *Language Assessment in Practice*, has no reference to validity in its index. It is clearly possible to discuss good practice in developing tests and justify-

ing testing, both in great depth and with great insight, without relying upon this problematic word.

Conclusion

We believe that a widespread professional consensus over how best to use the word 'validity' would be a good thing. Unfortunately, despite nearly 100 years of definitional debate:

(1) there is still no such consensus; in fact,
(2) we seem further now from consensus than ever before, and
(3) the nature and extent of disagreement is extreme, not trivial.

We have argued that the best way to use the word 'validity' cannot be determined on the basis of logical analysis alone. The fundamental question is whether the consequences which result from using the word in one way are manifestly better than the consequences which result from using the word in any other. Moreover, there are some good consequential arguments in favour of each of the opposing perspectives, which makes it very hard to decide between them, and which ensures that these debates are even more emotionally charged than they might otherwise be. We see no clear prospect of reaching a widespread professional consensus over a precise technical definition. If we are right, then what alternatives present themselves?

Clearly, we could just ignore the problem, and put up with the barrier to effective communication that a lack of consensus necessarily establishes. We could beseech future generations of graduate students just to work much harder to make sense of the conflicting representations of validity which they find in their textbooks. We could agree to live with widespread confusion over the meaning of the most important word of our lexicon, and suffer the inevitable consequences. Or, before resigning ourselves to this fate, we could explore alternatives that might lie outside of the box.

One leftfield suggestion is to stop trying to eliminate ambiguity over the word 'validity' and instead to embrace this ambiguity, as a community. This would require us to reach consensus *not* to use validity as a technical term, but only as a lay term. Another leftfield suggestion is to agree to retire the word 'validity' entirely, given irreconcilable differences of opinion over its proper application. This would require us to take seriously the possibility that our technical lexicon is sufficiently rich to allow us to retire the word with minimal negative impact. Given the centrality of 'validity' to present-day testing discourse, these suggestions may sound faintly ridiculous. They invite us to consider a radically new discourse, either with or without the most hallowed term in our current lexicon. Yet, the semantic anarchism of the current situation is equally ridiculous. Radical times may require radical action.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. The *Standards* were published originally as *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association, American Educational Research Association, & National Council on Measurements Used in Education [APA, AERA, & NCMUE], 1954); and most recently as *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). A chapter entitled *Validity* has appeared in all six editions of the *Standards*; containing a list of validity standards, prefaced by an extensive explanation of the validity concept. The definitions of validity which have been elaborated in these chapters have differed substantially over the years, despite the joint committees responsible for their production striving tirelessly to reach consensus. This is not simply a matter of the official definition changing over time, which is true. It is also a matter of validity being extremely consensus-resistant at any particular point in time. Describing their experience of producing the most recent edition, Plake and Wise (2014) observed that one 'ongoing tension was differing perspectives on validity theory' (p. 8).
2. For analysis of the second, see Borsboom, Cramer, Keivit, Scholten, and Franic (2009), Borsboom (2012), Newton (2012a, 2012b), Markus and Borsboom (2013), Borsboom and Markus (2013), Kane (2013a, 2013b), Newton and Shaw (2014).
3. The last two editions seem to have moved some way towards a more liberal position, although it is debateable quite how far they have moved in this direction (Newton, 2012a).
4. Even the new edition of the *Standards* – in which the authors appear to have exerted considerable effort to refer consistently throughout to the 'validity of test score interpretations for the intended use(s)' – sometimes slips into talk of 'test validity' (p. 49), 'valid measure' (p. 19), 'valid measurement' (p. 52), 'validity … of results' (p. 56), 'score validity' (p. 59), 'test validity' (p. 49), 'validity, reliability, and fairness of intended uses' (p. 77), 'validity of the classification procedure' (p. 30), 'validity and fairness of those decisions and practices' (p. 139), and so on.
5. Despite the widespread use of 'health' as a catchall term, certain organisations have, on occasion, felt a need to articulate a more precise technical definition. Perhaps the best known is the definition provided by the World Health Organisation (WHO) in 1948: 'Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.' (World Health Organisation, 2014). In arbitrating the matter, thus, WHO clarified that health could not simply be reduced to physical issues. However, its use of the word 'complete' set an extremely high bar for the concept, particularly given its insistence that governments have a responsibility for the health of their peoples. Subsequently, there has been considerable debate over the precise technical definition of the term, which continues to the present day (e.g. Bircher & Kuruvilla, 2014; Freeman, 2014; Huber et al., 2011). With this in mind, there is certainly something to be said for not seeking a precise technical definition, if one can be avoided.
6. In Newton and Shaw (2013), we may have muddied the waters by appearing to suggest that we could resolve a century of debate over the meaning of the word 'validity' simply by using the word 'quality' instead. This was disingenuous in two respects. First, by appearing to prioritise technical quality, we may have appeared to downplay social value. Second, by appearing to *substitute* the label 'quality' for the label 'validity', we may have appeared to trivialise the problem; surely, to borrow a metaphor from Shakespeare, a rose by any other name would smell as sweet? Our point was simply that the idea of technical quality already has a long pedigree within the testing community; e.g. 'Through application of these standards, tests have attained a high degree of quality and usefulness' (APA, AERA, & NCMUE, 1954, p. 1); 'The *Standards* is intended for professionals who specify, develop, or select tests and for those who interpret, or evaluate the technical quality of, test results' (AERA, APA, & NCME, 2014, p. 1). The idea of 'technical quality' helpfully orients interlocutors to a particular kind of analytical perspective, which can be unpacked using many other concepts that already have a long pedigree within the testing community: potential to measure; potential to improve decision-making; potential to bring about certain impacts; and so on. The problem, as we see it, is the desire to set *any* word on a pedestal as high as the one on which 'validity' currently sits. We certainly do not recommend that 'validity' should be dethroned and a new label

installed. We are simply suggesting that our lexicon already has sufficient words with which to discuss the full range of design and evaluation issues, without the need to rely upon 'validity'.

Notes on contributors

Paul E. Newton is a Research Chair at Ofqual, the Office of Qualifications and Examinations Regulation for England. His research focuses primarily upon issues related to the evaluation of large-scale educational assessment systems, and he is particularly interested in theories of validity for educational and psychological measurement, past and present. He has published on a range of assessment topics, including validity, comparability, assessment purposes, reliability and the public understanding of measurement inaccuracy.

Stuart D. Shaw is a principal research officer at Cambridge International Examinations (CIE). He is particularly interested in demonstrating how Cambridge Assessment in general and CIE in particular are seeking to meet the demands of validity in their assessments. He has a wide range of publications in English second language assessment and educational research journals.

Paul and Stuart co-authored the book *Validity in Educational and Psychological Assessment* (SAGE, 2014).

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2), 1–38.

Association of Language Testers in Europe. (2001). *Principles of good practice for ALTE examinations: Revised draft, October 2001*. Retrieved September 2014, from http://www.alte.org/attachments/files/good_practice.pdf

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Baker, E. L. (2013). The chimera of validity. *Teachers College Record, 115*(9), 1–26.

Bircher, J., & Kuruvilla, S. (2014). Defining health by addressing individual, social, and environmental determinants: New opportunities for health care and public health. *Journal of Public Health Policy, 35*, 363–386.

Borsboom, D. (2005). *Measuring the mind*. Cambridge: Cambridge University Press.

Borsboom, D. (2012). Whose consensus is it anyway? Scientific versus legalistic conceptions of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*, 38–41.

Borsboom, D., Cramer, A. O. J., Keivit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135–170). Charlotte, NC: Information Age.

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement, 50*, 110–114.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 85–116). New York, NY: Cambridge University Press.

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

Brennan, R. L. (2013). Commentary on "validating the interpretations and uses of test scores". *Journal of Educational Measurement, 50*, 74–83.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium-XIV. *Journal of Educational Psychology, 12*, 271–275.

Buckingham, B. R., McCall, W. A., Otis, A. S., Rugg, H. O., Trabue, M. R., & Courtis, S. A. (1921). Report of the standardization committee. *Journal of Educational Research, 4*, 78–80.

Cambridge Assessment. (2009). *The Cambridge approach: Principles for designing, administering and evaluating assessment*. Cambridge: Cambridge Assessment.

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

Chatterji, M. (2013). Insights, emerging taxonomies, and theories of action towards improving validity. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability an equity* (pp. 273–308). Bingley: Emerald Group.

Cizek, G. J. (2010). Error of measurement: Validity and the place of consequences. *NCME Newsletter, 18*, 4–5.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*, 31–43.

Crocker, L. (1997). Editorial: The great validity debate. *Educational Measurement: Issues and Practice, 16*, 4–4, 34.

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

Cronbach, L. J. (1976). Equity in selection – Where psychometrics and political philosophy meet. *Journal of Educational Measurement, 13*, 31–41.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Cronbach, L. J., Yalow, E., & Schaeffer, G. (1980). A mathematical structure for analyzing fairness in selection. *Personnel Psychology, 33*, 693–704.

Dunnette, M. D. (1992). It was nice to be there: Construct validity then and now. *Human Performance, 5*, 157–169.

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement, 34*, 39–48.

Finkelstein, L. (2009). Widely-defined measurement – An analysis of challenges. *Measurement, 42*, 1270–1277.

Forster, M. (2010). Wittgenstein on family resemblance concepts. In A. Ahmed (Ed.), *Wittgenstein's philosophical investigations* (pp. 66–87). Cambridge: Cambridge University Press.

Forte Fast, E., & Hebbler, S. with ASR-CAS Joint Study Group on Validity in Accountability Systems (2004). *A framework for examining validity in state accountability systems*. Washington, DC: Council of Chief State School Officers.

Freeman, P. (2014). What do we mean when we use the word *health*? *Journal of Public Health Policy, 35*, 357–362.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*, 21–28.

Gallie, W. B. (1956). Essentially contested concepts. *Proceedings of the Aristotelian Society, 56*, 167–198.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*, 456–462.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17*, 16–19, 34.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions* (2nd ed.). New York, NY: Routledge.

Haertel, E. (2013). Getting the help we need. *Journal of Educational Measurement, 50*, 84–90.

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology, 19*, 451–473.

Huber, M., Knottnerus, J. A., Green, L., van der Horst, H., Jadad, A. R., Kromhout, D., … Smid, H. (2011). How should we define health? *British Medical Journal, 343*, 235–237.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.
Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.
Kane, M. T. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50*, 115–122.
Kline, P. (1998). *The new psychometrics: Science, psychology and measurement*. London: Routledge.
Lees-Haley, P. R. (1996). Alice in validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51*, 981–983.
Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1–16.
Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*, 14–16.
Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*, 28–30.
Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age.
Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448.
Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research, XL*, 109–126.
Mameli, M., & Bateson, P. (2006). Innateness and the sciences. *Biology and Philosophy, 21*, 155–188.
Mameli, M., & Bateson, P. (2011). An evaluation of the concept of innateness. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*, 436–443.
Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8*, 435–461.
Maraun, M. D., Slaney, K. L., & Gabriel, S. M. (2009). The Augustinian methodological family of psychology. *New Ideas in Psychology, 27*, 148–162.
Marcoulides, G. A. (2004). Conceptual debates in evaluating measurement procedures. *Measurement: Interdisciplinary Research and Perspectives, 2*, 182–184.
Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.
Maul, A. (2013). On the ontology of psychological attributes. *Theory and Psychology, 23*, 752–769.
McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.
Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*, 16–18.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027.
Messick, S. (1989a). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Washington, DC: American Council on Education.
Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5–11.
Michell, J. (1999). *Measurement in psychology.* Cambridge: Cambridge University Press.
Michell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 111–133). Charlotte, NC: Information Age.
Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson Education.
Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229–258.
Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14*, 5–13.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*, 6–12.
Newton, P. E. (2012a). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1–2), 1–29.
Newton, P. E. (2012b). Questioning the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*, 110–122.
Newton, P. E., & Shaw, S. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*, 301–319.
Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. London: Sage.
Norris, S. P. (1995). Measurement by tests and consequences of test use. In *Philosophy of Education* (pp. 303–306). Urbana, IL: Philosophy of Education Society.
Odell, C. W. (1928). *A glossary of three hundred terms used in educational measurement and research*. Urbana: University of Illinois.
Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal* (2nd ed.). Boston, MA: Pearson Education.
Penzes, W. B. (n.d.). *Time line for the definition of the meter. A contribution by the Precision Engineering Division of the Manufacturing Engineering Lab at the National Institute of Standards and Technology*. Retrieved September 2014, from http://www.glb.nist.gov/pml/div683/museum-timeline.cfm
Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME *Standards for Educational and Psychological Testing*? *Educational Measurement: Issues and Practice, 33*, 4–12.
Popham, W. J. (1997). Consequential validity: Right concern – Wrong concept. *Educational Measurement: Issues and Practice, 16*, 9–13.
Pressey, S. L. (1920). Suggestions looking toward a fundamental revision of current statistical procedure, as applied to tests. *Psychological Review, 27*, 466–472.
Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*, 13–16.
Ruch, G. M. (1924). *The improvement of the written examination*. Chicago, IL: Scott, Foresman.
Sackett, P. R. (1998). Performance assessment in education and professional certification: Lessons for personnel selection. In H. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 113–129). Mahwah, NJ: Lawrence Erlbaum Associates.
Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont, CA: Wadsworth.
Scriven, M. (2002). Assessing six assumptions in assessment. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 255–275). Mahwah, NJ: Lawrence Erlbaum Associates.
Shepard, L. A. (1990). Readiness testing in local school districts: An analysis of backdoor policies. *Journal of Education Policy, 5*, 159–179.
Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.
Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5–8.
Simmons, S. J. (1989). Health: A concept analysis. *International Journal of Nursing Studies, 26*, 155–161.
Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*, 477–481.
Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565–578.
Tenopyr, M. L. (1998). Measure me not: The test taker's new bill of rights. In H. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 17–22). Mahwah, NJ: Lawrence Erlbaum Associates.

Thurstone, L. L. (1931). *The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems*. Ann Arbor, MI: Edwards Brothers.

Urbina, S. (2004). *Essentials of psychological testing*. Hoboken, NJ: Wiley.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 75–107).Hillsdale, NJ: Lawrence Erlbaum Associates.

World Health Organisation. (2014). *WHO definition of health* [Webpage]. Retrieved September 2014, from http://www.who.int/about/definition/en/print.html