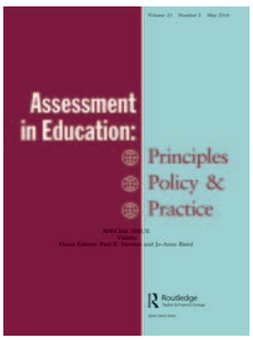

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

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

# Evaluating test validity: reprise and progress

Lorrie A. Shepard

**To cite this article:** Lorrie A. Shepard (2016) Evaluating test validity: reprise and progress, Assessment in Education: Principles, Policy & Practice, 23:2, 268-280, DOI: 10.1080/0969594X.2016.1141168

**To link to this article:** http://dx.doi.org/10.1080/0969594X.2016.1141168

*Assessment in Education: Principles, Policy & Practice*, 2016
Vol. 23, No. 2, 268–280, http://dx.doi.org/10.1080/0969594X.2016.1141168





## REFLECTIVE OVERVIEW

# Evaluating test validity: reprise and progress

Lorrie A. Shepard*

*School of Education, University of Colorado Boulder, Colorado, CO, USA*

(*Received 22 October 2015; accepted 8 January 2016*)

The AERA, APA, NCME Standards define validity as 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests'. A century of disagreement about validity does not mean that there has not been substantial progress. This consensus definition brings together interpretations and use so that it is one idea, not a sequence of steps. Just as test design is framed by a particular context of use, so too must validation research focus on the adequacy of tests for specific purposes. The consensus definition also carries forward major reforms in validity theory begun in the 1970s that rejected separate types of validity evidence for different types of tests, e.g. content validity for achievement tests and predictive correlations for employment tests. When the current definition refers to both 'evidence and theory' the Standards are requiring not just that a test be well designed based on theory but that evidence be collected to verify that the test device is working as intended. Having taught policy-makers, citizens, and the courts to use the word validity, especially in high-stakes applications, we cannot after the fact substitute a more limited, technical definition of validity. An official definition provides clarity even for those who disagree, because it serves as a touchstone and obliges them to acknowledge when they are departing from it.

I was invited to provide a reflective commentary in response to the focal papers in this special validity issue of *Assessment in Education*. The authors hold different views on the definition of validity and were selected by the editors to highlight important points of disagreement. To be as clear and focused as possible, I have chosen to defend what Newton and Shaw (2016) characterise as the 'liberal'[1] position in the debate about whether test validity should be narrowly and technically defined or should include critical aspects of test use. To engage this debate, I focus primarily on the positions advanced by Newton and Shaw (2016) and Cizek (2016) in this volume. I also acknowledge the contribution of Borsboom, Mellenbergh, and van Heerden (2004) because their simple definition of validity is frequently cited by those holding to a more 'conservative' position.

My response is structured around the following main points.

(1) Definitional simplicity should not be an end in itself. Given that validation research is essentially theory testing, we should not be surprised at its complexity.
(2) Having taught policy-makers, citizens and the courts to use the word validity, especially in high-stakes applications, we cannot after the fact substitute

*Email: lorrie.shepard@colorado.edu

a more limited, technical definition of validity.

(3) Test design is always framed by the intended purpose of a test, and similarly validation research should be organised to evaluate validity for particular uses.

(4) Messick's matrix was a mistake because it led to false dichotomies between aspects of validity that are inherently entwined.

(5) Complaints that the current definition is too encompassing to be practical are clearly addressed by Kane's (1992, 2006) validity argument approach, which helps set priorities by identifying those claims that are most critical to investigate – including evidence regarding both intended and unintended consequences.

In crafting this essay, I elected to focus on the dimensions of the conservative vs. liberal debate rather than engage with each paper in isolation. With apologies, I necessarily give short shrift to the papers by Moss (2016) and Markus (2016). Moss (2016) offers an expanded theory of validity and accompanying research agenda regarding data use in school settings. Regarding her paper, I should clarify only that her exposition is quite expansive; in my view, it is possible to examine both intended and unintended test effects within a validity framework without going so far as to examine affordances beyond the test that also affect data use. Markus's (2016) paper is itself a commentary on existing definitions and theories of validity. Rather than take sides regarding particular points of contention, he offers an analytic schema for translating claims between contrasting vocabularies and hence a means to deepen our understanding of specific issues.

## Embracing complexity

Newton and Shaw (2016) argue that a precise definition of validity is needed to remove ambiguity and arrive at greater professional consensus. In somewhat the same vein, Borsboom et al. (2004) reject complex conceptions of validity extant in the literature and offer, instead, an alternative that is 'simple, clear, and workable' (p. 1061). Yet, nowhere is it written that simplicity per se will ensure the wisdom of scientific inquiry. A wise saying, attributed to Einstein, is that 'things should be as simple as possible, but no simpler'.

Although they extoll simplicity, Borsboom et al.'s (2004) chief argument has to do with the need for a strong theory that accounts for examinee response processes instead of relying on atheoretical, correlational approaches to establish validity. According to Borsboom et al. (2004), 'a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure' (p. 1061). I agree with their stance, after Embretson (1983), requiring that first consideration be given to content representation. It is reasonable to say that validity is partly (but only partly) built into a test by having a strong theoretical understanding of the construct to be measured and how it is expected to be manifest in test scores. I also appreciate their disdain for bootstrapping constructs out of correlation matrices.

I disagree with Borsboom et al.'s (2004) treatment of causation, however, but not because we don't want there to be a theory-of-the-test-construct that explains variation in test outcomes. Rather, I disagree because their entire exposition fails to consider how variation in test scores is to be interpreted and used in the face of

multiple causes. Although they do not say they require it, their definition of validity only works for deterministic cause, where all of test score variation is accounted for by variation in the attribute. It cannot work for probabilistic causes. In the social sciences, except for tautologies, causes aren't deterministic. To use their example of the attribute *intelligence*, wouldn't we dispute the validity of the test if scores were causally affected both by intelligence and also by opportunities to practice similar questions in advance of the test? Shouldn't the conflating of causes be part of the validity evaluation? Of course, one could design quasi-experimental studies to control for multiple causes, but Borsboom et al. (2004) do not acknowledge a need for such controls to say that a test is functioning as intended.

Borsboom et al. (2004) do admit, however, that there is a price to be paid for simplicity. In their own words, important ideas such as test bias are explicitly left out and would have to be investigated separately (and called something other than validity evidence).

> The validity concept proposed here has been stripped of all excess baggage. The benefit is that this lends substantial clarity and force to the concept, but the price is that it covers less ground. For instance, when one claims validity, one is not thereby claiming reliability, predictive adequacy, or absence of bias. (p. 1070)

In my view, they have made the concept too simple, and I argue in the next section that such a move – to leave out things like bias and predictive accuracy – would play false with public audiences who have quite different expectations when we assure them of a test's validity.

While shorthand definitions may be useful, especially for lay audiences, an understanding of validity and validation research cannot be reduced to concise, stipulative definitions, as if these were merely vocabulary terms. It's a good thing, not a detriment that experts recognise the complexity of these endeavours and devote whole chapters, in the *Standards* and in the *Educational Measurement* (Brennan, 2006) bible, to validity rather than offering only short, declarative definitions. While I disagree with Borsboom et al.'s (2004) claim that attributes can be gotten at independent of our logical and empirical ways of knowing, I agree with the point they make to distinguish between validity, as a property, and validation, as the activity researchers undertake to evaluate validity. They say, 'Validation is more like theory testing' (p. 1063). I agree and note, therefore, that validation research is necessarily complex. Complexity does not mean, however, that validation is a nebulous swirl. As taken up in a later section, using a validity argument to frame an investigation allows for effective tailoring and prioritisation of the evidence needed for particular testing applications.

## Why the term 'validity' and not test 'quality'

Borsboom et al. (2004) propose that a term like 'overall quality' could be used instead of validity to refer to a more complete set of test properties in need of evaluation, such as 'reliability, predictive adequacy, or absence of bias' (p. 1070). Newton and Shaw (2016) similarly like this idea of using terms other than validity to 'clarify core concepts' (p. 9) for evaluating educational testing. The problem, however, with inventing new terms to replace the encompassing version of validity is that this valued and believed-to-be-scientific thing is now presumed by broad audiences of researchers, practitioners and the public at large. The Wikipedia *test validity* entry,

for example, last updated in July 2015, offers the familiar saying that, 'Test validity is the extent to which a test accurately measures what it purports to measure', but then goes on in the next breath to say that, 'In fields of psychological testing and educational testing, "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests"' (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). On one recent date, this Wikipedia page had been viewed 2484 times in the previous 30 days and 7515 times in the previous 90 days.

A Google search of the phrase 'reliable, valid, and fair' produced 56,300 hits in .31 s. 'Reliability, validity, and fairness' produced even more, with 62,600 hits. The entry 'test validity' identified 37.5 million documents. Of course, it would be silly to claim that these multitudes of users hold common definitions of the terms. It is not foolish, however, to suggest that, when seeking a formal definition, students and scholars most often turn to the consensus definition from the 1999 *Standards* as evidenced by the Wikipedia entry. It might also be a fair bet that the vast majority of these professional and lay users of the terms believe, when they think about validity, something like 'a test is adequate for its purpose' and not just that some of the variation in test scores is caused by the named attribute.

Testing experts in psychology and education developed these terms of art and used them as a short-hand to communicate with other professionals, courts and judges, parents and other test users. Although physicists and chemists are free to coin esoteric terms without having to be bound by public understandings, measurement scientists do not have this luxury, especially when we started with a common word and built our specialised meanings on top of it. The vocabulary of social science is embedded in institutional and legal contexts and laden with associated connotations. Unlike the vocabulary of the natural sciences, 'validity' is a term used in the context of evaluating teachers and students and controlling access to higher education, employment, gifted education and so forth. It would be an inappropriate bait-and-switch tactic to deploy a narrower definition in these contexts in which validity is the more complex and decision-directed idea developed in the broader institutional–legal context.

It might be helpful to recall that the reshaping of the definition of validity, beginning in the 1970s and continuing through to the 1985 and 1999 versions of the *Standards,* came about in the context of the civil rights movement because tests were being used to sort and segregate in harmful ways that could not be defended as valid. In 1971, in Griggs v. Duke Power Co, the US Supreme Court ruled against the use of intelligence tests to select employees for higher level jobs because the test had a discriminatory impact on blacks and the company lacked evidence that the test was related to job performance. Referring to Congressional intent in the Civil Rights Act of 1964, under which the suit was brought, Justice Burger's opinion for the unanimous court ended with the statement that, 'What Congress has commanded is that any tests used must measure the person for the job and not the person in the abstract' (at 436). My citing of Griggs v. Duke Power Co is consistent with the history of validity theory provided by Kane (2016) and illustrates specifically how the courts caused experts to rethink the definition of validity. The quotation from Justice Burger, for example, meant that to be adequate for the intended use, the test could not just be shown to measure intelligence but had to be shown to measure intelligence relevant to the job.

Similarly in Larry P. v. Riles, in 1972, the District Court in California issued an injunction against the use of IQ tests to place black students in classes for the educable mentally retarded (EMR). The school district claimed that black children with IQ scores below 75 could not benefit from instruction offered in regular education classes and would be better off in EMR classes. Unconvinced, the court was persuaded, instead, by the claims of racial bias because of far greater racial disproportion in California, where IQ tests were the primary means for making placement decisions, compared to the rates in other states where criteria such as achievement tests and teacher evaluations were used. In addition, Judge Peckham accepted the plaintiffs' claims that erroneous placement in EMR class caused irreparable injury because of a dead-end curriculum focused more on social skills and grooming instead of academic skills and because of the shame and ridicule experienced by EMR students. More is said in the last section of the paper about the need to focus a validity investigation on the test purpose and intended benefit, which in this example was the claim that the test would match assessed attributes of children with an appropriate treatment. The test being used might have been a valid measure of intelligence in some other context, but in this context, it was not valid for determining whether children did or did not have sufficient intelligence to benefit from instruction in the regular classroom.

The definition of validity provided in the new 2014 *Standards* is as follows. (It differs only slightly from the more widely accessed 1999 document.)

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests*. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. (AERA, APA, & NCME, 2014, p. 11)

This is referred to as the consensus definition because it was arrived at by a joint committee of experts following two stages of extensive review and commentary by dozens of professional associations, credentialing organisations and testing companies. The fact that the 1999 and 2014 definitions are so similar also speaks to shared understandings over time by a large majority of testing experts. The consensus process does not mean that this definition is absolutely right, nor is it fixed for all time. It does mean that this is the professionally defensible definition to be shared with non-experts, and it is the definition that experts and testing companies should acknowledge, if they decide to depart from it. Rather than portraying the field as hopelessly in disarray, because some experts disagree, I would argue that the consensus definition provides a clear and well-organised framework for orderly debate. Instead of a simple and unanimous definition, this is the agreed-upon working definition of a complex idea and research agenda. Importantly, if you decide to leave out critical elements from the *Standards* definition, you are obliged to provide an explanation.

## Putting test use at the centre of validity investigations

It is a cardinal rule of our field that test design depends on intended purpose. Why would we say that use matters for test design but not for evaluating test validity? Validation research is in many ways an evaluation of whether the test design was

adequate for its purpose. When a test is to be developed, it is not sufficient to say only that the test will measure reading. In addition to specifying the age or grade level for which the reading test is intended, it is also critically important to distinguish among uses such as measuring the effectiveness of educational systems, evaluating growth from fall to spring, diagnosing learning disabilities or planning short-term instructional interventions. For some uses, it is critical to include material that is well below and well above the target grade level. When comparing educational systems, the test content must be 'curriculum fair' across participating states or countries. For placement decisions, greater measurement precision is needed at the cut-score. If intended for instructional interventions, then subtest scores should be sufficiently reliable to warrant differentiated decisions about which skills require further support. This is not test use tacked on at the end of test design. It is a clear understanding of test use that shapes test design throughout. Of course, tests may also be designed for multiple purposes; and similarly validity should be evaluated for each of those purposes.

A helpful example, illustrating how use shapes design, is provided by the distinctions drawn in *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001) between classroom and large-scale assessments. In a coherent assessment system the two levels of assessment would share a common model of learning, but would nonetheless be designed quite differently. The *Knowing What Students Know* authors invoked the idea of trade-offs in test design proposed by Cronbach and Gleser (1965) to address the tensions between *fidelity* and *bandwidth*. Some testing purposes require much greater measurement precision and substantive depth, while other purposes call for greater breadth at the expense of precision for individual test takers. Classroom teachers need detailed information about individual students directly tied to specific units of study, whereas state-level policy leaders want comprehensive information across all of the curricular standards intended for a grade level. We can go further and note that contextual features of assessment tasks that add meaning and interest for instructional purposes can lead to those same items being rejected when designing a large-scale assessment. Think of a mathematics problem set involving the weight of snow on a roof, which would be considered biased if the test is to be administered for large-scale, comparative purposes in both Colorado and Florida. The same issues – following from intended purpose – that shape test design must in turn guide test validity studies.

Sireci (2016) summarises the development and evolution of validity theory over the past 100 years, emphasising the attention to test use as early as 1920. Sireci cites Rulon's (1946) well known statement that, 'we cannot label a test valid or not valid except for some purpose' (p. 290). In an earlier debate, I cited Curton's (1951) validity chapter from the very first edition of *Educational Measurement*:

> The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for the third. (p. 621)

Those who argue that validity is only about test score interpretation and not about the validity of subsequent decisions and actions must either take the position that test results will sit on a shelf and not be used, or they admit that there are other questions of adequacy and legitimacy but they want to call these something other than validity questions.

As noted previously, it is misleading, when speaking to audiences of researchers or the public, to claim that evidence of validity has been established while purposely neglecting intended contexts of use. One might argue that tests used only for research and not for practical applications could get by with only evidence of valid interpretations. But, in fact, research contexts also require evidence of validity for the specific use as much as for any applied purpose. A reading test would not be valid for determining the relative benefits of phonics-only reading instruction, for example, if it were called a reading test but included only phonics and vocabulary subtests and not measures of comprehension. We would say that such a test was not a valid measure of reading in that case. I am grateful to philosopher of education, Ken Howe (personal communication) for noting that part of our shared understanding of reading is 'its positive normative valence, which links score interpretation with subsequent consequences. It is the nature of terms such as reading to be intrinsically related to action, lest they have no interest or use'.

As with other authors, classified by Newton and Shaw (2014) as 'conservatives', Cizek (2016) wants to use a different word for aspects of validity that go beyond the narrowly technical. He prefers the term justification of test use rather than validation. Cizek (2016) insists that 'validation of an intended score inference and justification of a specific test use – are not only separable, they *cannot* be combined (emphasis in original)' (p. 5). He gives two examples that are unhelpful in quite different respects. The first is an absurd example where we are asked whether a biology end-of-course test should be used for awarding high school diplomas. The second is a more realistic example. 'Do these ACT/SAT scores measure high school preparation for success in college?' and a second question, which he says is outside the validity scope, 'Should these ACT/SAT scores be used for college admission decisions' (p. 6). It should be noted that in the second example, the construct intended to be measured 'high school preparation for success in college' has been defined so broadly that it includes the intended use and the implied criterion performance. So, to respond to the score inference question, we would be well on our way to a full blown validity investigation. If, for example, students with high ACT/SAT scores still needed a remedial writing course, that would be evidence that the test lacked validity for the intended inference and use.

To elaborate his distinction between validation and justification, Cizek (2016) provides a table that lists examples of the types of evidence that would be needed to evaluate the validity of score meaning. Because these examples are all stated generically, it may not be obvious to the reader that the validation evidence, which the table suggests only pertain to score meaning, will in fact depend on test use. For example, Cizek lists 'content/curricular alignment studies' as a source of evidence based on test content that would be needed for validating score meaning. I would argue, however, that a measurement specialist embarking on such a study would need to orient toward two different curriculum frameworks depending upon whether validity was being judged in terms of the test's adequacy for assessing district progress toward new state standards or to award or withhold high school diplomas for students receiving instruction under the old standards. The difference between the current and hoped-for curriculum (as represented by new standards) is a legitimate part of a validity investigation and should be attended to when claiming validity for the second purpose. Test purpose will also make a difference when implementing many of the empirical examples on the validation/score-inference side of Cizek's table. As noted earlier, 'correlations among subtests' can be high for some

large-scale purposes but not for individual tests that claim diagnostic utility. 'Correlations with criterion variables' imply some particular test use and presumably the studied populations would not be some hit-and-miss sample but would be representative of the intended context of use. 'Investigations of mean differences for relevant groups' would be a necessary part of a validity study when use of a test is intended to improve learning, either as the treatment variable (in the case of formative assessment) or as a placement test designed to better match instruction with learning needs. Identifying the groups to be compared in any given investigation would depend on the logic model for the intended test use.

Let me be clear, not every possible aspect of test use is automatically a part of an appropriate validity evaluation. How much a test costs is not part of validity nor is the question of whether a school district can afford to buy enough computers for online administration. It is a question of validity, however, if students reading and writing abilities are obscured by lack of familiarity with the computer interface, just as it has always been a validity problem when kindergarteners lack the fine motor coordination to bubble in answer choices. When I say that test use should be at the centre of validity investigations, I am calling for the same attention that is given to intended use and intended effects in test design. We include certain content, carry out field tests and refine with particular populations and uses in mind. Similarly, test publishers advertise for specific purposes and promised benefits. Validation research then necessarily focuses on these goals or valued ends.



## Fact-value Distinctions and the Mistake of Messick's Matrix

In a review in 1993 and again in 1997, I quarrelled with Messick's (1989) matrix, not because I disagreed with his ideas, but because his famous 2 × 2 table and segmented discussion invited the kinds of unwarranted separations that we see are still causing misinterpretations today (Shepard, 1993, 1997). The whole thrust of Messick's monumental validity chapter in the third edition of *Educational Measurement* was to argue that validity is a unitary concept. As noted in numerous historical accounts, a unitary and encompassing definition of validity became an increasingly important shared understanding during the 1970s and 1980s, driven by the heat of controversies regarding IQ testing. The new unitary concept organised under the framework of construct validation represented a rejection of past content-only or correlation-only types of validity.

It is unfortunate then, that in elaborating upon a unitary conception that brought together logical and empirical sources of evidence for all manner of tests, Messick made the mistake of offering a new set of *facets*. He cautioned that they were entwined and that construct validity resides in each of the cells, but nonetheless, the labels of his rows and columns have invited separations that Messick and other validity theorists in those decades wanted to foreclose. The left and right columns of Messick's table were labelled 'Test Interpretation' and 'Test Use'. In the previous section, using Cizek's examples of sources of evidence, I attempted to illustrate how the sources of evidence selected for evaluating test score interpretation will, in fact, depend on test use, not just the name of the test construct. Messick's two columns were intended to deepen our thinking about each aspect, but these aspects or facets cannot be addressed independently. While Cizek's arguments are based on a left–right divide of Messick's table, Newton and Shaw (2016) appear to argue for a

top–bottom division between the rows that Messick had labelled the 'Evidential Basis' and 'Consequential Basis' of test validity, which are similarly comingled.

Newton and Shaw's binary, separating facts from values and scientific evaluation from ethical evaluation, is not an analysis that can be defended given current understandings in the philosophy of science (Howe, 2009; Putnam, 2002). Newton and Shaw characterise the liberal position by saying that it 'fully embraces the idea that it is insufficient, if not irresponsible, to evaluate tests from a purely scientific or technical perspective' (p. 4). While I have been content in this paper to distinguish my view from the narrowly technical 'traditionalist' or 'conservative' position, it cannot be argued that theirs is the 'scientific' perspective. To act as if value choices and ethical decisions are outside of science is to ignore the value-laden nature of the scientific process involved in every aspect of test development and validity evaluation, which Messick pointed out at length. How constructs are named slants their meaning. How they are defined and operationalised, who gets to review for relevance, what populations are selected for standardisation, which group differences are taken out and which are left in, whether there is bias in the criterion measure, whether a test creates bigger black–white differences than can be observed in performance on the job – in all of these cases, value considerations are entirely embedded in the steps of doing science. They cannot be disentangled.



**Validity argument and attention to consequences**

Again because of the matrix, Messick (1989) is sometimes cited as if he 'added' consideration of the social consequences of tests to the concept of validity, when in fact he merely elaborated and called our attention to a long-standing, fundamental aspect of validity studies that follows from attention to test use. What might be considered new in the 1980s, in response to concerns about test bias, was a re-centring of validity studies on intended effects. What did 'test use' mean if not a set of claims about how using a test was expected to lead to particular desired outcomes? And once attention was focused explicitly on intended effects of a testing programme, it followed inevitably that unintended effects should also be considered – one of many valuable lessons learned from re-conceiving validity research as programme evaluation (Cronbach, 1988). Thus an IQ test might be sufficiently 'valid' for use in a research study to evaluate the long-term effects of foetal alcohol syndrome on cognitive functioning but not be valid for placing children in special education.

A powerful case in point was provided by the National Research Council Panel on selection and placement of students for programmes for the mentally retarded led by Heller, Holtzman, and Messick (1982). The Heller et al. Panel had been convened to examine the causes and possible biases leading to overrepresentation of minority children and males in classes for mentally retarded students. The Panel redefined its charge, however, in a way that exemplifies the re-centring of contemporary validity theory on *the adequacy of a test for achieving its intended outcomes*. In the case of special education placements, the intended outcome was to provide more effective educational interventions – tailored to the student's needs – than would be available in the regular classroom. By asking the larger question as to why disproportion was a problem, the Panel brought into their analysis the bodies of research showing the negative effects of labelling and the poor quality of instruction in segregated special education classrooms. To be valid, they said, an assessment should address a child's functional needs that could be linked to effective interventions. 'Thus, assessments

can be judged in terms of their utility in moving the child toward appropriate educational goals (p. 99)'. The Panel reprised the well-known science emphasising that IQ tests measure current cognitive functioning rather than an inborn trait, but they also made this conclusion beside the point if current functioning could not be matched to effective treatments, given that achieving an effective match was what was being claimed. Measures of reading comprehension or adaptive behaviour had the potential to be more useful, but significantly these measures too could not be claimed to be valid for making placements, if placements were shown to be ineffective.

Following Cronbach (1988), Kane (1992), and others, the 1999 *Standards* adopted *validity argument* as the framework for organising and integrating validity evidence. 'Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use (p. 9)'. Beyond test use and testing consequences, per se, an argument-based approach to validity furthered the refinement of validity theory in two important ways: (1) it required that the theory of the test be made explicit, with underlying assumptions revealed, and (2) it helped to establish priorities as to which evaluative questions were the most central to investigate. For example, in Kane's (1992) example of an algebra placement test used to assign college students either to a calculus or remedial algebra course, the argument for the test depended on several assumptions, including the assumption that algebraic skills are genuinely prerequisite and used in the calculus course, that the placement test represented those skills well and that the remedial course was effective in teaching the target algebraic skills. What might not be so obvious was a final assumption that high-scoring students placed directly in calculus would *not* also benefit, i.e. their performance in calculus would not improve, if they too had received the remedial algebra treatment. In other words, differential placement also had to be shown to be effective. Note that Kane (1992, 2001, 2006) has consistently identified two layers to the argument approach, the interpretive argument (which lays out the logic model and substantive claims) and the validity argument (which entails the gathering and analysis of evidence to support the plausibility of the interpretive argument). As a short-hand, the *Standards* and others have combined both layers in what is called for in a validity argument approach, and I use that simplification here as well.

Complaints that the current definition of validity is too encompassing to be practical are addressed by using a validity argument approach to lay out the theory of the test and identify the particular claims about benefits that are most critical to investigate. It is also not necessary to pursue all possible unintended consequences. Cronbach (1988) advised that, 'Validation advances on a broad front, what we learn about one test bears on others' (p. 4). Especially it is the case that likely negative effects become known about an entire class of tests and should be used to focus investigations of unintended consequences. For example, a large research literature on high-stakes testing has documented the redirection of instructional time to tested subjects and to imitations of testing formats that produce test-score inflation (Haladyna, Nolen, & Haas, 1991; Herman, 2008; Jacob, 2005; US Congress Office of Technology Assessment, 1992). It follows, then, that when interim assessments are developed and marketed, with the intention of improving learning by enabling early intervention, the claim of improved learning requires validity evidence. And, knowing that practicing test formats might lead to spurious results, an outcome measure is needed that addresses this validity concern.

Cizek (2016) especially complained that the ambitious validity theory carried forward in the most recent version of the *Standards* is so onerous that 'an integrated evaluative summary of the evidence' has never been accomplished. I believe that the difficulty here is that Cizek is looking for a single publication to provide this synthesis rather than a programme of research. In a 1997 review, I provided several examples of programmes of validation research, noting especially studies undertaken by the National Research Council. These included the Heller et al. (1982) study of IQ tests used for special education placements, cited above, and the Panel report addressing the use of the General Aptitude Test Battery to make employee referral and selection decisions (Hartigan & Wigdor, 1989). Other examples included the body of research informing the validity of the Scholastic Aptitude Test (SAT) for college selection decisions (so long as it is not used for strict, top-down selection) and the programme of research by Shepard and Graue (1993) and others pointing to the invalidity of the Gesell School Readiness Screening Test to keep low-scoring children out of kindergarten. A more up-to-date example is the extensive set of studies undertaken by the NAEP Validity Studies Panel.

I understand that we do not have at hand a plethora of complete, tied-in-a-bow, validation syntheses carrying through all of the steps of the validity argument for particular tests in their contexts of use. It does not follow, however, that because comprehensive programmes of research are infrequent that validity argument is the wrong framing for whatever validation research is undertaken. Especially we should not revert to narrow, technical content or correlational analyses and then act as if there is robust scientific evidence of the adequacy of the test's performance. This is the type of practice that Cronbach (1989) had typified as 'raking together of miscellaneous correlations' (p. 155) in contrast to systematic checks organised by an interpretive argument. If limited evidence is to be collected, i.e. only one or two studies instead of a full validity argument investigation, then those one or two studies should address the most significant claims being made rather than collecting only the data that are most readily at hand.

## Concluding remarks

Disagreements about the definition of validity, as identified by Newton and Shaw (2016), are not necessarily a problem in a scientific field, so long as we are clear about the nature of the disagreement and track how differences in our conceptions lead in turn to differences in methods and findings. In the face of less than perfect agreement, I have argued here that the consensus definition as represented in the 2014 *Standards* is the best working definition, in part simply because it was arrived at by a consensus process, involving not only an expert committee but feedback from the field and multi-organisational review and approval. Having an official definition provides clarity even for those who disagree because it serves as a touchstone and obliges them to acknowledge how and why they are departing from it.

One hundred years of disagreement about validity as characterised by Newton and Shaw (2016), does not mean that there has not been at the same time substantial progress. Importantly the consensus definition brings together interpretations and use so that it is one idea, not a sequence of steps. As I have emphasised here, test design always begins not just with the name of the attribute or construct to be measured but also with an understanding of how that evidence will be used. We don't completely design a reading test and then find out after the fact that it was intended

to be used for adults, not third graders. Therefore, just as test design is framed by a particular context of use, so too must validation research focus on the adequacy of tests for specific purposes. The consensus definition also carries forward major reforms in validity theory begun in the 1970s that rejected separate types of validity evidence for different types of tests, e.g. content validity for achievement tests and predictive correlations for employment tests. When the current definition refers to both 'evidence and theory' (AERA, APA, & NCME, 2014, p. 11), the *Standards* are requiring not just that a test be well designed based on theory but that evidence be collected to verify that the test device is working as intended.



## Disclosure statement

No potential conflict of interest was reported by the author.

## Notes on contributor

Lorrie A Shepard is Dean of the School of Education and Distinguished Professor at the University of Colorado Boulder. Her research, focused on the use and misuse of tests in educational settings, has addressed the identification of learning disabilities, readiness screening for kindergarten, grade retention, teacher testing, effects of high-stakes accountability testing, and most recently the use of classroom assessment to support teaching and learning. Her technical work has contributed to validity theory, standard setting, and statistical models for detecting test bias.

## Note

1. Strictly speaking my position is more like what Newton and Shaw (2014) label as 'moderate'. However, their distinction between liberal and moderate positions requires a separation of scientific and ethical considerations, which I argue later in this paper is not possible.

## References

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

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

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

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Civil Rights Act of 1964, Pub. L. No. 88-352, 78 Stat. 241. (July 7, 1964).

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy and Practice*, *23*, doi: 10.1080/0969594X.2015.1063479.

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. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 147–171). Urbana, IL: University of Illinois Press.

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

Curton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.
Griggs et al. v. Duke Power Co., 401 U.S. 424 (1971).
Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*, 2–7.
Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the general aptitude test battery*. Washington, DC: National Academy Press.
Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity*. Washington, DC: National Academy Press.
Herman, J. L. (2008). Accountability and assessment: Is public interest in K-12 education being served? In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 211–231). New York, NY: Routledge.
Howe, K. R. (2009). Positivist dogmas, rhetoric, and the education science question. *Educational Researcher, 38*, 428–440.
Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*, 761–796.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.
Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger Publishers.
Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy and Practice, 23*, doi: 10.1080/0969594X.2015.1060192.
Larry P. v. Riles, 343 F. Supp. 1306 (N.D. Cal. 1972).
Markus, K. A. (2016). Alternative vocabularies in the test validity literature. *Assessment in Education: Principles, Policy and Practice, 23*, doi: 10.1080/0969594X.2015.1060191.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.
Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy and Practice, 23*, doi: 10.1080/0969594X.2015.1072085.
Newton, P. E., & Shaw, S. D. (2014, April 2–6). *Do we need to use the term 'validity'?* Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.
Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy and Practice, 23*, doi: 10.1080/0969594X.2015.1037241.
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.
Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16*, 290–296.
Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 19, pp. 405–450). Washington, DC: American Educational Research Association.
Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(5–8), 13.
Shepard, L. A. & Graue, M. E. (1993). The morass of school readiness screening: Research on test use and test validity. *Handbook of Research on the Education of Young Children*, 293–305.
Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy and Practice, 23*, doi: 10.1080/0969594X.2015.1072084.
US Congress Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: US Government Printing Office.