Assessment in Education: Principles, Policy & Practice, 2015
http://dx.doi.org/10.1080/0969594X.2015.1063479

Routledge
Taylor & Francis Group

Validating test score meaning and defending test score use: different aims, different methods

Gregory J. Cizek*

School of Education, Program in Learning Sciences and Psychological Studies, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

(*Received 23 October 2014; accepted 15 June 2015*)

Advances in validity theory and alacrity in validation practice have suffered because the term *validity* has been used to refer to two incompatible concerns: (1) the degree of support for specified interpretations of test scores (i.e. intended score meaning) and (2) the degree of support for specified applications (i.e. intended test uses). This article provides a brief summary of current validity theory, explication of a critical flaw in the current conceptualisation of validity, and a framework that both accommodates and differentiates validation of test score inferences and justification of test use.

Keywords: validity; validation; justification

It is difficult to overstate the importance of measurement either in theory or in applied contexts, such as educational testing. As Cone and Foster have observed, 'Scholars commonly acknowledge that developments in all areas of science follow discoveries of appropriate measurement techniques', and they argue that 'measurement provides the foundation for all other scientific pursuits' (Cone & Foster, 1991, p. 653). Indeed, measurement concerns are routinely at the forefront of training, practice and research in the social sciences and are omnipresent in the development and evaluation of tests. The importance of measurement is heightened whenever tests results inform consequential decisions including, for example, when evaluating the effectiveness of interventions, when awarding credentials, licences or diplomas, and when making personnel decisions. Fundamentally, the central focus of measurement specialists is the quality of data yielded by tests. Test takers, test users and all consumers of test information benefit from this focus.

Broadly endorsed professional standards for tests have existed for over a half century, beginning with the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association, 1954) and spanning six editions to the current *Standards for Educational and Psychological Testing* (hereafter, *Standards*; American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME], 2014). Although the *Standards* have evolved, the primacy of one topic – validity – has been consistently affirmed. The current *Standards* describe

*Email: cizek@unc.edu

validity to be 'the most fundamental consideration in developing and evaluating tests' (AERA, APA, NCME, 2014, p. 11).

Its nominal primacy notwithstanding, the concept of validity and the practice of validation have languished needlessly. Following a brief introduction to contemporary validity theory, this article will describe a lingering flaw in the concept of validity and a comprehensive remedy to address the error will be presented.

Contemporary validity theory

Although divergent viewpoints regarding validity are presented in this special issue, there is actually broad consensus about several aspects of contemporary validity theory. For one, there is wide agreement that validity pertains to the intended inferences or interpretations made from test scores and not to the tests themselves (Cronbach, 1971; Messick, 1989).

For another, although a multitude of 'validities' has been coined (see Newton & Shaw, 2013), contemporary validity theorists largely reject the existence of diverse kinds of validities, in favour of a unitary concept of validity (Messick, 1989, 1995). The unitary view centres on the extent to which available evidence supports interpretations of scores to reflect standing with respect to a specified construct (Cizek, 2012; Messick, 1998).

Third, there is consensus that judgments about validity are not dichotomous (e.g. test scores are not valid or invalid) but are appropriately described along a continuum of evidentiary support for the intended score inferences (Zumbo, 2007). The process of validation involves gathering, synthesising and evaluating data that typically includes both confirming and disconfirming evidence.

Fourth, there is agreement that the validation process is not a one-time activity (Shepard, 1993). Instead, factors such as experience with administration and scoring of the instrument, changes in the intended test population, the availability of previously unknown or unavailable sources of evidence, replications of the initial validation effort, theoretical evolution of the construct itself, and many other factors can alter the original judgments. Thus, on-going reviews of the original support for any conclusions about validity are necessary.

Finally, there is agreement that the process of validation necessarily involves the application of values (Messick, 1975). For example, among the many junctures at which value judgments are made in the testing process, values are brought to bear when deciding which sources of validity evidence should be mined, the relevance of those sources, how they should be weighted, and the favourability/unfavourability of the evidence. Even the most comprehensive validation efforts can yield equivocal evidence that can lead equally qualified evaluators to different conclusions; those conclusions depend on beliefs, assumptions and values that affect perceptions of the validity evidence to be synthesised (Longino, 1990).

A fatal flaw in contemporary validity theory

Despite broad endorsement of the importance of validity and agreement on many tenets of modern validity theory, there are also areas of disagreement (see, for example, Borsboom, Craver, Kievit, Scholter, & Franic, 2009; Borsboom, Mellenbergh, & van Heerden, 2004; Hood, 2009; Lissitz & Samuelsen, 2007). A comprehensive

review of all points of disagreement is beyond the scope of this article. However, one area of concern is perhaps the most consequential – a flaw in the very definition of validity. Given its elevated status as 'the most fundamental consideration' in testing (AERA, APA, NCME, 2014, p. 11), it would seem that a clear, accessible, broadly accepted definition of validity would exist. It does not. Remarkably, most contemporary authors have avoided proffering a crisp definition of the term (see, for example, Elmore & Camilli, 2007; Kane, 2006; Wainer & Braun, 1988).

Perhaps the most familiar and oft-cited description of validity is that provided by Messick (1989) where the concept is defined as 'an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment' (p. 13, emphasis in original). A number of theorists have pointed out weaknesses in Messick's definition (see, e.g. Borsboom et al., 2004; Hood, 2009). However, perhaps the most fatal flaw – with both theoretical and practical consequences – is that validity is defined with an inherent internal contradiction. Even the most recent edition of the *Standards* perpetuates the conflation of score meaning and test use, indicating that validity 'refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (AERA, APA, NCME, 2014, p. 11).

As can be seen in Messick's (1989) and the *Standards* definitions, validity is defined as two very different things: namely, validity is defined as (a) the extent to which evidence supports the intended meaning of the test scores *and* (b) the extent to which the subsequent actions or consequences of using a test align with (implicit or explicit) values and intended uses.

As Cizek (2011, 2012) has demonstrated previously, these two endeavours – validation of an intended score inference and justification of a specific test use – are not only separable, they *cannot* be combined. Perhaps the most compelling evidence confirming this flaw is that no 'integrated evaluative judgment' of the type described by Messick (1989) has ever been produced. The envisioned synthesis of evidence bearing on the accuracy of test score inferences and evidence bearing on the appropriateness of test score use is neither logically nor practically possible.

Distinguishing between score meaning and test score use

For clarity, the two essential aims of defensible test development and use can be simplified as straightforward research questions:

Q1: What do these scores mean?
Q2: Should these test scores be used for X? (where X is a specific test use).

Practical examples of these questions are illustrated in the following pairs of interrogatories:

Q1: Do these End-of-Course scores reflect mastery of the high school biology content standards?
Q2: Should these biology End-of-Course test scores be used for awarding high school diplomas?

and,

Q1: Do these ACT/SAT scores measure high school preparation for success in college?
Q2: Should these ACT/SAT scores be used for college admission decisions?

Clearly the above pairs each capture very different – but equally important – questions regarding judgments about the intended score inference and the justification of a specific test use. A diverse array of empirical and logical rationales confirms that the two questions require distinct sources of evidence bearing on the differing purposes and that a single synthesis of evidence on both inference and use is not possible.

At a fundamental level, the two questions reflect different targets of inquiry. It has been fairly well established in other fields that 'judgment and justification are separate processes' (Haidt, 2012; Johnson-Laird & Wason, 1977). Differing justification processes can lead to potentially conflicting conclusions regarding an issue, question, claim or target of an investigation.

From a more practical perspective, any evidence gathered on one of the questions is non-compensatory with respect to the other, and evidence gathered is typically more relevant to one of the questions than the other. For example, strong content validity evidence for a mandated biology end-of-course test for high school graduation supports the inference that the test measures biology knowledge and skill; such evidence would be necessary to support the use of the test for making any interpretations related to competence in biology, but would not beyond that provide support for using the test as a basis for awarding high school diplomas. (That is, if it was determined that the test did *not* have strong content validity, it would be difficult to argue that it could be defensibly used in a system where content mastery ostensibly undergirded graduation decisions.) Thus, validation of score inferences is a necessary first step – but not a sufficient condition – for justification of a test use. Conversely, evidence that use of the biology test as a graduation requirement increased (or decreased) student persistence in high school would add no support for the claim that the test was well-aligned to the biology curriculum.

In general, the strongest possible evidence supporting any aspect of validity (e.g. evidence based on test content, response process, internal structure, relationships to other variables) says nothing about whether test scores should be used for any specific purpose; the strongest possible evidence that the use of a test has certain valued benefits or detrimental consequences says nothing about what the test scores mean. The meaning, interpretation, or inference based on the test result – that is, the *validity* of the test scores – is typically unaffected by actions based on the test scores, the uses of the test results or the consequences of those uses.

To be clear, there are rare occasions when data obtained following test use can cycle back to inform the claims made about the construct a test purports to measure. This occurs when post-testing evidence reveals mis- or under-specification of the construct. An illustration of this has been provided by Guion (1980) in his presidential address to the APA's Division of Industrial and Organisational Psychology. Guion described an experiment in which male and female participants were judged on their speed in packaging golf balls into cartons from an assembly line that was placed at a specified distance from the participants. A very short distance between participants and the assembly line advantaged females (who, on average, had shorter arms than males), compared to males who found the working conditions too cramped for rapid movement. The resulting consequence data – that is, the greater 'failure' rate for the

males – were evidence that the construct 'packing speed' had been mis-specified and a source of construct irrelevant variance (arm length) was affecting the test results. Importantly, the illustration also shows how consequences did not affect the accuracy of an inference – that is, that females *were*, on average, speedier packagers of golf balls under the specified conditions. It also clearly shows how evidence obtained after a test has been administered can be valuable in identifying aspects of the test design or testing conditions that are not consonant with the intended inference about the construct. Unfortunately, it is not typically this kind of construct-relevant evidence that is meant when 'consequential validity' is referenced; rather, policy implications or other social consequences of testing are invoked as bearing on the validity of a test when they have no relationship whatsoever to the meaning of the test scores.

There is one other way in which the validity of test score inferences and the defensibility of an intended test use are related. Although decisions about how to use the test scores for some purpose (or even to use the test at all) must be made, it is perhaps obvious that those decisions necessarily presume that the test scores can be interpreted as valid representations of the constructs they are intended to measure. However, it should be equally clear that even substantial evidence regarding score validity does not dictate what decisions about test use should be. Mehrens has argued as follows:

> [S]uppose an adult male has an elevated PSA reading. The issue of whether this is an accurate indicant of prostate cancer is separable from the consequences of whatever treatment may follow. To confound them seems unwise. (Mehrens, 1997, p. 16)

Unwise and illogical. Overall, strong evidence supporting the intended meaning of scores yielded by a test is a necessary but insufficient condition for justifying any specific use of the test. It would be professionally reckless to use test scores for any given purpose when the very meaning of scores produced by the test is unsupported or unclear.

A revised framework for defensible testing

Figure 1 provides a revised framework for defensible testing. The framework shows how each of the two major concerns in sound test development and use is

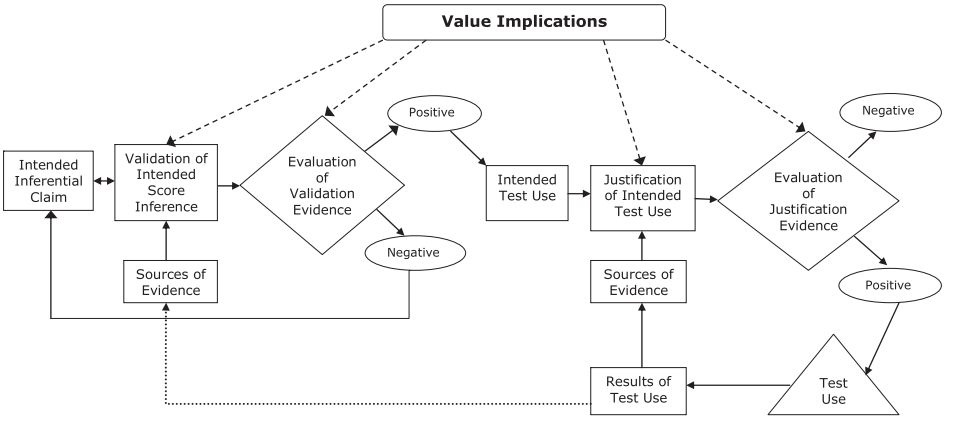

Figure 1. Processes of validating test score inferences and justifying test score use.

distinguished. Validation of an intended score inference (the left half of the figure) is presented as a separate process from justification of a proposed test use (the right half).

Examining the figure from left to right, the process of the test validation and justification can be seen as progressing in a logical sequence, with validation work to gather evidence in support of the intended score meaning occurring primarily during test development and evaluation (although, as indicated previously, validation is an on-going process). Efforts to gather evidence to justify an intended test use occurring later, primarily after score meaning has been supported (although, as indicated previously, justification efforts can begin when a test use is contemplated and certainly include information gathered on the consequences of test use). At the top of the figure, the overarching label, Value Implications, highlights the fact that values come into play at all junctures in the testing process.

Figure 1 shows that defensible test development and use begins with a clear statement of the Intended Inferential Claim. This statement guides the validation effort and gathering of evidence which is then evaluated with respect to the support provided for the claim. The bidirectional arrow between the Intended Inferential Claim and Validation of Intended Score Inference reflects a recursiveness in which the gathering and evaluation of validity evidence prompts re-examination and refinement of the intended inferential claim, which in turn suggests alternative validation strategies and sources of evidence.

The process illustrated in Figure 1 then shows that an integrative Evaluation of the Validation Evidence results in an overall judgment of the extent to which the evidence supports the claim (Positive) or is disconfirming (Negative). If the validity investigation yields adequate evidence in support of the Intended Inferential Claim, effort then shifts primarily towards clear specification of the Intended Test Use and Justification of the Intended Test Use. Like evaluation of the validation evidence, Evaluation of the Justification Evidence can yield either affirmation (Positive) or rejection (Negative) of the intended use. In the final phase of the process, actual operational Test Use, additional information is generated that bears on justification of the use (e.g. anticipated benefits and consequences of testing are observed) and, in some circumstances, can provide confirming or disconfirming evidence vis a vis score meaning.

Finally, two important caveats should be noted. First, as has been recommended elsewhere (AERA, APA, NCME, 2014), the processes illustrated in Figure 1 should be followed for each intended score interpretation and for each intended test use. Second, the concerns of validation and justification often interact. Evidence of the validity of score inferences is a necessary but insufficient condition for recommending or sustaining a justification for test use. Validity evidence is an essential part and precursor of the justification for the use of a test – but only a part – and one that may carry greater or lesser weight in deliberations concerning test use. As Borsboom and Mellenbergh have stated in the context of tests used for selection, placement or with the goal of bringing about changes in society at large, ‘validity may play an important role in these processes, but it cannot by itself justify them’ (Borsboom & Mellenbergh, 2007, p. 109).

Evidence for defensible testing

A common element in Figure 1 is ‘Sources of Evidence,’ which appears for both Validation of Intended Score Inference and for Justification of Intended Test Use.

However, as has been argued elsewhere (Cizek, 2012) and is a central thesis of this article, the differing purposes of validation and justification necessarily require the collection, synthesis, and evaluation of differing sources of evidence. Defensible testing is supported when there is evidence that test scores can be interpreted confidently as intended *and* there is evidence that intended uses of the scores are justified. One implication of the revised framework presented in Figure 1 is that a reconsideration and reconfiguration of these sources of evidence is also in order.

With regard to sources of evidence for score meaning – that is, *validity* evidence – there are long-standing and well-developed guidelines drawing, primarily, on psychometric traditions. The current *Standards* (AERA, APA, NCME, 2014) list five source of possible validity evidence. Of those, four seem appropriate: Evidence Based on Test Content (p. 14), Evidence Based on Response Processes (p. 15), Evidence Based on Internal Structure (p. 16) and Evidence Based on Relations to Other Variables (p. 16). The fifth – Evidence for Validity and Consequences of Testing (p. 19) – is the outlier. Of the five, it is the only source of evidence bears directly on the question of whether it is justifiable to *use* a test, not on the meaning of the test score.

Obviously, a critical need for the future is to develop other sources of evidence for justification of test use to the same refined state as in the current menu of psychometric sources of evidence for validation of score interpretations. Whereas the psychometric tradition has provided a solid foundation for developing sources of evidence bearing on validation of intended score meaning, it seems ill-suited as a framework for developing sources of evidence bearing on justification of test use.

Some potentially appropriate frameworks for justification of test use exist; one example can be found in the theory and methods of programme evaluation – a paradigm that comprises much of what would ordinarily be considered essential when contemplating the use of a test for a given purpose (see, e.g. Donaldson, Christie, & Mark, 2009; Patton, 2008; Shadish, Cook, & Leviton, 1991; Stake, 2004). Programme evaluators typically investigate precisely the kinds of concerns that arise in the context of test use; namely, the merit, worth, or significance of the object of evaluation (Scriven, 1995); they are typically concerned with aspects such as utility, feasibility, propriety and accuracy (Yarbrough, Shulha, Hopson, & Caruthers, 2011). And, the field of programme evaluation has developed systematic strategies for investigating the issues that often underlie justification of test use, such as needs assessment, cost-benefit approaches, cost-utility analyses and cost-effectiveness investigations. Perhaps most importantly given the often-contested potential uses of test scores, the field of programme evaluation has long recognised and incorporated the necessity to identify stakeholders, the realities of differential negotiation acumen and rhetorical skill among stakeholders, the interplay of politics in evaluation decision-making, and contested notions of desirable outcomes. In addition, the concepts related to fairness in testing explicated by Camilli (2006) are relevant to the justification of an intended test use.

Using the lens of programme evaluation, justification of test use begins with a clear articulation of its intended use. (This is analogous to the first step in validation, which comprises clear articulation of the intended score meaning.) As in the practice of programme evaluation, next steps would include: identification of stakeholders and affected audiences; specification of explicit evaluation questions to be addressed (depending on chosen evaluation method); gathering and evaluating credible data

that bear on those questions (Centers for Disease Control & Prevention, 1999); and documenting the evaluation process.

Evidence bearing on justification of an intended test score use

A complete exploration of sources of evidence for justification of test use is beyond the scope of this article. Indeed, one aim of this article was to stimulate the beginnings of theory development and practical guidance related to the appropriate sources of evidence for justification of a test use. Nonetheless, some possible sources of evidence come readily to mind. Table 1 represents a first attempt to unify several aspects of defensible testing. The first column in the table has four entries; for the most part, each entry reflects the familiar sources of evidence for validation of intended score inferences. Some elaboration on these four sources of validity evidence will be presented shortly. The second column provides some basic examples of each of these sources of validity evidence.

A beginning exploration of evidentiary sources related to test use is also provided in Table 1. Possible sources of evidence related to justification of intended test use are listed in the third column of the table; the fourth column provides examples of each source of evidence. The table is purposefully formatted using parallelism in presentation to highlight a unified approach to defensible testing.

The sources of evidence for Justification of Intended Test Use shown in Table 1 include: (1) Evidence Based on Consequences of Testing; (2) Evidence Based on Costs of Testing; (3) Evidence Based on Alternatives to Testing and (4) Evidence Based on Fundamental Fairness. Some examples of each of these potential sources of evidence are suggested in the last column of Table 1. It should be noted that, like the sources of evidence for validating an intended score meaning, not all of the possible sources of evidence for justifying a test use would ordinarily be pursued. The particular constellation of evidentiary sources is necessarily linked to the specific concerns or questions of interest. And, as suggested earlier in this article, if more than one test use is contemplated, a justification effort for each use would be required.

Evidence bearing on validation of an intended score interpretation

To a great extent, the sources of evidence that can be gathered to provide support for an intended score inference have been well articulated elsewhere. As indicated previously, the *Standards* (AERA, APA, NCME, 2014) chapter on validity offers good guidance on this effort, listing potential sources of support for the meaning of a test score. An examination of the sources of evidence for validating score meaning listed in Table 1 (see column 1) reveals that much of the conventional wisdom has been maintained. For example, the first two sources of evidence for validating score meaning are identical to those listed in the *Standards* for nearly 30 years: Evidence Based on Test Content and Evidence Based on Response Processes. However, Table 1 also suggests some small but important differences compared to a list derived directly from the existing *Standards* for sources of validity evidence.

Three differences are noteworthy. First, as argued previously, Evidence Based on Consequences of Testing has been reclassified under the more relevant endeavour – that of justifying test use. Although it appears as the first entry in the column of potential Sources of Evidence for Justifying Test Use, it should be noted that it is

Table 1. Sources and examples of evidence for validation of score meaning and justification of test use.

Sources of evidence for validating score meaning	Examples	Sources of evidence for justifying test use	Examples
Evidence based on test content	• Grounding in relevant theoretical dimensions or relationships • Job analyses • Content/curricular alignment studies	*Evidence based on consequences of testing*	• Evaluation of anticipated benefits • Consideration of negative consequences • Consideration of false-positive, false-negative rates
Evidence based on response processes	• Cognitive labs • Think-aloud protocols • Cognitive mapping	*Evidence based on costs of testing*	• Overall cost of testing • Cost-benefit evaluation • Consideration of opportunity costs
Evidence based on hypothesised relationships among variables	*Internal* • Coefficient alpha, KR-20 • Confirmatory factor analysis • Correlations among subscores *External* • Correlations with criterion variables; convergent and discriminant analyses • Investigations of mean differences for relevant groups (e.g. treated/ untreated; males/females.) • Multi-trait, multi-method analyses	*Evidence based on alternatives to testing*	• Evaluation of relative value of alternative testing methods, formats, or procedures • Evaluation of non-test options to accomplish intended goals
Evidence based on test development and administration procedures	• Item/task generation procedures • Bias/sensitivity reviews • Test administration and scoring procedures • Test security procedures	*Evidence based on fundamental fairness*	• Evaluation of stakeholder inclusion • Investigation of opportunity to learn • Provision of due notice • Examination of differential impact across relevant subpopulations

but one of four equally valuable sources of evidence, depending on the specific intended test use.

Second, a new source of evidence for validating intended score interpretations is shown as the last entry in the first column: Evidence Based on Test Development and Administration Processes. This category recognises sources of evidence that promote accurate score interpretations, but which heretofore have been undervalued and remained unacknowledged in the current version of the *Standards*. These sources include procedures such as judgmental bias and sensitivity reviews, adequate implementation timelines, protocols designed to ensure test security/test score integrity, deployment of user familiarisation applications and for computer-based testing, and quality assurance procedures for test scoring.

Finally, the sources of validity evidence called 'Evidence Based on Internal Structure' and 'Evidence Based on Relations to Other Variables' in the *Standards* (AERA, APA, NCME, 2014, p. 16) have been subsumed under the single heading of 'Evidence Based on Hypothesised Relationships among Variables.' This suggested change is intended to recognise the familiar principle that 'reliability is a necessary but insufficient condition for validity,' to progress further towards a unified framework for defensible testing, and to redress an artificial distinction in the current *Standards*. A brief explanation of the artificial distinction follows.

As just indicated, the current *Standards* include a source of validity evidence labelled 'Evidence Based on Internal Structure.' Perhaps the most widely recognised quantification of internal structure is Cronbach's Alpha, a now-ubiquitous *reliability* index first published by Lee Cronbach in his 1951 article 'Coefficient Alpha and the Internal Structure of Tests.' Although coefficient alpha and many other such indices have been developed primarily to provide information about score reliability, the internal structure of a test clearly bears on the validity of score interpretations, so restricting their usefulness only to the dependability of scores would be mistaken. Thus, it seems clear that such sources of reliability evidence also bear on the intended meaning of score interpretations, that is validity.

Moreover – and bearing on the rationale for combining 'Evidence Based on Internal Structure' and 'Evidence Based on Relations to Other Variables' as listed in the current *Standards*, under the single heading of 'Evidence Based on Hypothesised Relationships among Variables' – is the fact that the separate sources represent a distinction with an ignorable difference: both essentially examine relationships among variables. The current – and largely trivial – distinction is that investigations of internal structure (e.g. alpha, KR-20, and factor analysis.) examine variables that are internal to the test itself, whereas investigations of relations to other variables typically focus on relationships between test scores and variables external to the test. That is, the variables typically studied under the current heading of 'internal structure' are the items that comprise a measure and test takers' responses to those items; the variables typically studied under the current heading of 'other variables' are test takers' responses to the test items and their responses on other measures.

It seems clear that, in both cases, it is relationships among variables that are being examined – the only difference being the nominal classification of where the variable was obtained. The singular category of 'Evidence Based on Hypothesised Relationships among Variables' is proposed to not only provide a more coherent, unified framework for examining relationships among variables, but the notion of *hypothesised* relationships is added to connote that any such investigations should be theory based. That is, when examining the internal structure of a test, one should

proceed based on a theoretically grounded working hypothesis of what the internal dimensions of a test are intended to be; likewise, the examination and interpretation of relationships between test scores and variables external to the test under study should be guided by theoretically grounded positions about how those variables should be related.

Conclusions and next steps

A reconceptualization of validity is long overdue. Regrettably, the flaw in validity theory is maintained in the just-released version of the *Standards*.

Spanning now more than four decades, from when the notion of so-called consequential validity was first introduced until today, the measurement field has wrestled with the attempts to incorporate attention to two fundamental measurement concerns – score meaning and test use – awkwardly subsumed under the single heading of 'Validity' (cf., Brennan, 2006; Tenopyr, 1975). Despite professional admonitions that test developers and users should produce integrated evaluations of evidence bearing on the different foci of score meaning and score use, there is no evidence that the desired outcome – an integrated evaluative summary of the evidence – has ever successfully been accomplished. This is almost certainly because the sources of information that might be mined bear on distinctly different questions and a synthesis of that evidence is not possible. It is appropriate to abandon the pretence and advice that evidence bearing on validation of score meaning and that bearing on justification of test use can be combined in any meaningful way, and to instead provide test developers and test users with rigorous strategies for gathering, synthesising and evaluating evidence on each of these important aspects of defensible testing.

It seems clear that aims and processes of evidence gathering in support of answering the question 'What does this score mean?' can – and indeed must – be distinguished from evidence gathering in response to the question 'Should these scores be used for … ?' It seems equally clear that a comprehensive approach to defensible testing must give equal attention to both concerns and cannot continue to perpetuate a blending and blurring of evidence bearing on each into a single 'integrated' synthesis called validity. Finally, systematic, rigorous and differentiated professional guidelines – such as those enumerated in the *Standards* – are needed in support of evidence specification, gathering and evaluation for the equally important and distinct endeavours of validating test score interpretation(s) and justifying test scores use(s).

Although standard advice related to the systematic gathering of evidence related to test score meaning exists, a comprehensive approach to defensible testing would also provide standard advice related to rigorous evidence gathering in support of justifying test use. As can be inferred from Figure 1, at a conceptual level, the processes of validating score meaning and justifying test use are quite similar, although the aims of each endeavour differ substantially. Thus, the specific sources of evidence that are germane to addressing the two questions differ, as do the methods used for gathering and evaluating that evidence. The discipline of programme evaluation might provide one fruitful avenue for beginning to develop rigorous guidelines for evidence gathering and evaluation related to justifying test score us; other theoretically relevant and practically useful paradigms should be investigated and developed.

Commenting on the conflation of test score meaning and test score use in the current *Standards*, Kane has noted that in developing [the *Standards*], the organisations have put the requirement for evaluation of proposed interpretations and uses under the heading of validity. We can change that and put some or all of these issues under some other heading … but if we do so, we will have to reformulate much of the standard advice provided to test developers and test users (Kane, 2009, p. 62).

Yes, we will. And we should: Why would we hesitate to do so? All of the models, textbooks and other materials placing the earth at the centre of the universe needed to be revised when Copernicus described the inherent contradictions in then-current conceptualizations and proposed a helio-centric perspective. If it is not plausible to argue that validating score meaning and justifying test use are the same thing, then it seems counterproductive not to revise the assessment models, textbooks and other materials as well.

Validity theory has advanced appreciably and continues to be an evolving concept. Modern psychometrics has moved far from the notion that 'a test is valid for anything with which it correlates' (Guilford, 1946, p. 429) to a more sophisticated paradigm with many, broadly accepted, fundamentals. The shift to the unitary view of validity – indeed, perhaps all such substantial scientific reconceptualizations – are characteristically not universally embraced initially, but gain increased acceptance as the intractabilities of the dominant paradigm become more widely recognised (Kuhn, 1962).

Modern validity theory must continue to evolve. Failure to do so would perpetuate perhaps the greatest injustice in modern testing. The flaw in current validity theory is more than simply a semantic inconvenience; it explains not only lingering controversy about several aspects of validity but also – and of greatest concern – validation in practice that is too often anaemic (Ebel, 1961) and justification in practice that too often reflects a Machiavellian approach wherein simply the loudest, most persuasive, and most powerful or well-funded voices determine the legitimate uses of tests. Decisions about the appropriate use of a test should depend more on the evidence supporting the intended use than the rhetorical force of any entity with a pecuniary or political interest in its implementation. For the greater good of those who are the consumers of test data, and to make progress towards a more comprehensive approach to defensible testing practice, we must begin to pursue the potential for more systematic, rigorous, transparent and democratic justification efforts that rival those that have been developed for validation of score meaning. To the extent that the concept of validity is more sharply defined and more coherent and that systematic justification efforts are stimulated, a revised conceptualisation that differentiates between validation of score meaning and justification of test score use can help foster the goals of facilitating more complete and searching validation practice. Such efforts also have potential to enhance the quality and utility of test results and to enabling those who develop and use tests to improve the outcomes for their clients, students, organisations and others that are the ultimate beneficiaries of high-quality test information.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Gregory J. Cizek is a professor of Educational Measurement and Evaluation at the University of North Carolina-Chapel Hill (USA) where he teaches courses in psychometrics, assessment, statistics, research methods and program evaluation. His scholarly interests include standard setting, validity, test security and testing policy. He is a contributor to the Handbook of Test Development (2006, 2015); editor of the *Handbook of educational policy* (1999) and *Setting performance standards* (2001, 2012); co-editor of the *Handbook of formative assessment* (2010, with H. Andrade); and author of *Cheating on tests: How to do It, detect it, and prevent it* (1999), *Detecting and preventing classroom cheating* (2003), *Addressing test anxiety in a high-stakes environment* (with S. Burg, 2005) and *Standard setting: A practitioner's guide* (with M. Bunch, 2007). He provides expert consultation at the state and national level on testing programmes and policy, including service as a member of the National Assessment Governing Board which oversees the National Assessment of Educational Progress (NAEP). He has worked in leadership positions in the American Educational Research Association (AERA) and is past president of the National Council on Measurement in Education (NCME). Cizek has managed national licensure and certification testing programs and worked on test development for a statewide testing programme. He began his career as an elementary school teacher, and has served as an elected member of a local board of education.

References

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

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

Borsboom, D., Craver, A. O. J., Kievit, R. A., Scholter, 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. 135–172). Charlotte, NC: Information Age.

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). Cambridge: Cambridge University Press.

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

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: Praeger.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: Praeger.

Centers for Disease Control and Prevention. (1999, September 17). *Framework for program evaluation in public health*. (Morbidity and Mortality Weekly Report (RR No. 11)). pp. 1–58. Atlanta, GA: Author.

Cizek, G. J. (2011, April). *Error of measurement: Reconsidering validity theory and the place of consequences*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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

Cone, J. D., & Foster, S. L. (1991). Training in measurement: Always the bridesmaid. *American Psychologist, 46*, 653–654.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

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

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.). (2009). *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: Sage.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16*, 640–647.
Elmore, P. B., & Camilli, G. (Eds.). (2007). Special issue on validity. *Educational Researcher, 36*, 185–241.
Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427–439.
Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology, 11*, 385–398.
Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Random House.
Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory and Psychology, 19*, 451–473.
Johnson-Laird, P. N., & Wason, P. C. (1977). *Thinking: Readings in cognitive science*. Oxford: Cambridge University Press.
Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.
Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39–64). Charlotte, NC: Information Age.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.
Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448.
Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry.* Princeton, NJ: Princeton University Press.
Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*, 16–19.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.
Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35–44.
Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*, 301–319.
Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Newbury Park, CA: Sage.
Scriven, M. (1995). The logic of evaluation and evaluation practice. *New Directions for Evaluation,* 1995, 49–70.
Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practices*. Newbury Park, CA: Sage.
Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.
Stake, R. E. (2004). *Standards-based and responsive evaluation*. Thousand Oaks, CA: Sage.
Tenopyr, M. L. (1975, July). *Content-construct confusion*. Paper presented at the Content Validity II Conference, Bowling Green State University, Bowling Green, OH.
Wainer, H., & Braun, H. I. (Eds.). (1988). *Test validity.* Hillsdale, NJ: Erlbaum.
Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.
Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (pp. 45–79). Amsterdam: Elsevier Science.