Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 1 | Issue 1 Article 3

2020

De-“Constructing” Test Validation

Stephen G. Sireci

De-"Constructing" Test Validation

Stephen G. Sireci

University of Massachusetts Amherst

Abstract

Construct validity theory presents the most comprehensive description of "validity" as it pertains to educational and psychological testing. The term "construct validity" was introduced in 1954 in the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association [APA], 1954), and subsequently elucidated by two members of the 1954 committee — Cronbach and Meehl (1955). Construct validity theory has had enormous impact on the theoretical descriptions of validity, but it was not explicitly supported by the last two versions of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 1999, 2014). In this article I trace some of the history of the debate regarding the importance of construct validity theory for test validation, identify the essential elements of construct validity theory that are critical for validating the use of a test for a particular purpose, and propose a framework for test validation that focuses on test use, rather than test construct. This "de-constructed" approach involves four steps: (a) clearly articulating testing purposes, (b) identifying potential negative consequences of test use, (c) crossing test purposes and potential misuses with the five sources of validity evidence listed in the AERA et al. (2014) *Standards for Educational and Psychological Testing*, and (d) prioritizing the sources of validity evidence needed to build a sound validity argument that focuses on test use and consequences. The goals of deconstructed validation are to embrace the major tenets involved in construct validity theory by using them to develop a coherent and comprehensive validity argument that is comprehensible to psychometricians, court justices, policy makers, and the general public; and is consistent with the AERA et al. (2014) *Standards*.

Keywords

Construct validity; testing standards; validity

Validity theory in educational and psychological testing can be traced back over 100 years (e.g., Thorndike, 1904) and was concomitant with the emergence of standardized testing in the United States and Europe (Sireci, 2009). Over the past century, the concept of validity has been widely debated and several theories and sets of terminology have been proposed (Newton & Shaw, 2013). The first attempt to form a consensus view of validity was begun by the American Psychological Association (APA), in 1952 when they published *Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal* (APA, 1952). This proposal for professional guidelines on test development, use, and evaluation led to the first formal version of standards for educational and psychological testing — *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954) — a joint effort of APA, the American Educational Research Association (AERA), and the National Council on Measurements Used in Education (NCME[1]). These *Recommendations* set the stage for the beginning of a consensus within the measurement community regarding what validity was and how to go about the process of test validation. As part of this consensus view, it introduced a new concept — *construct validity*.

Two members of the committee that produced the 1954

[1]In 1961, the National Council on Measurements Used in Education changed its name to the National Council on Measurement in Education, which remains its current name (see NCME.org).

CONTACT: Stephen G. Sireci. sireci@acad.umass.edu. University of Massachusetts Amherst. 813 North Pleasant St. Amherst, MA 01003.

Recommendations, Lee J. Cronbach and Paul Meehl, published a subsequent paper to explain "construct validity" more completely (Cronbach & Meehl, 1955). This paper became one of the most seminal works in validity theory, and represented the beginning of what many consider to be a consensus view that all validity in educational and psychological testing is construct validity. However, like many consensus views in validity theory, construct validity theory eventually lost its eminence, although many of its major tenets remain important for test validation.

In this article, I trace a small portion of the history of construct validity theory to explain why many of the most respected validity theorists in the history of testing described validity in terms of a unitary concept centered on construct validity. Next, drawing from the seven versions of the *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*) that appeared between 1952 and 2014, I explain why this concept has not been emphasized in test validation efforts. In the third section, I focus on the modern definition of validity described in the last two versions of the *Standards* (AERA et al., 1999, 2014), and illustrate how the spirit of construct validity theory is embodied in a validation approach that uses the AERA et al. (1999, 2014) *Standards*' five sources of validity evidence as the framework for test validation. The purpose of this article is to tour the historical literature on validity theory to explain the key concepts that have endured and remain important in 20th-century test validation.

Brief Description of "Traditional" Validity Terminology

Before focusing on the history of construct validity, it is important to acknowledge the unfortunate perseverance of the notion that there are different types of validity, which seems to permeate many introductory textbooks in psychology and education. In Table 1, I present the different validity terms that were used in the seven versions of the AERA, APA, and NCME *Standards*. As can been seen in Table 1, validity terminology evolved from "categories" of validity, to "types" of validity, to "aspects" of validity, back to "categories," and finally to "sources of validity evidence," where they have remained for more than two decades. Although it is interesting that "content" is the only term that endured throughout the seven-decade history of the *Standards*, what many consider to be the three "traditional forms" of validity are: construct validity, content validity, and criterion-related validity. I will describe construct validity in the next section by tracing its history. Here, I provide only very brief descriptions of the traditional notions of content and criterion-related validity (see Kane 2006, 2013 for more comprehensive descriptions).

Content validity refers to the degree to which the content of a test represents the proficiencies targeted by the test, as well as the degree to which that content is consistent with the testing purposes. Content validity is an obvious prerequisite for educational and credentialing tests because such tests need to demonstrate "alignment" with the targeted curriculum or job domain (Crocker, 2003; Martone & Sireci, 2009; Sireci, 1998; Sireci & Faulkner-Bond, 2014). Content validity is typically evaluated by using subject matter experts to review test items and provide judgments of their relevance and representativeness with respect to the domain tested.

Criterion-related validity refers to statistical evaluation of the relationships of test scores to other variables of importance (i.e., criteria) for evaluating test score interpretations. Criterion-related validity studies are often partitioned into those that focus on *predictive validity*, where test scores are evaluated with respect to a criterion they are designed to predict (e.g., the degree to which college admissions tests scores predict college grade point average), and *concurrent validity*, where test scores are evaluated with respect to their relationships with other variables gathered at the same point in time (e.g., correlations between 5th-grade math test scores and students' grades in their 5th-grade math class). In the last two versions of the *Standards* (AERA et al., 1999, 2014) the notion of criterion-related validity is embodied as "validity evidence based on relations to other variables."

In a later section of this article, I revisit the modern notions of these different sources of validity evidence and their importance for validation. Next, I provide a brief introduction to construct validity theory.

1 Construct Validity: A Brief History

To understand construct validity, we must start with an understanding of the philosophical concept of a "construct." Cronbach and Meehl (1955) defined a *construct* as,

> some postulated attribute of people, assumed to be reflected in test performance...We expect a person at any time to possess or not possess a qualitative attribute...or to possess some degree of a quantitative attribute...(p. 283).

In describing construct validity, they stated,

> construct validation is involved whenever a test is to be interpreted as a measure of some attribute or

Table 1. Summary of Validity Terminology Used in Current and Prior Versions of the *Standards*

Publication	Validity Terminology
Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal (APA, 1952)	*Categories of validity*: predictive, status, content, congruent
Technical recommendations for psychological tests and diagnostic techniques (APA, 1954)	*Types of validity*: construct, concurrent, predictive, content
Standards for educational and psychological tests and manuals (AERA et al., 1966)	*Types*: criterion-related, construct-related, content-related
Standards for educational and psychological tests (AERA et al., 1974)	*Aspects of validity*: criterion-related, construct-related, content-related
Standards for educational and psychological testing (AERA et al., 1985)	*Categories of validity*: criterion-related, construct-related, content-related
Standards for educational and psychological testing (AERA et al., 1999)	*Sources of validity evidence*: test content, response processes, internal structure, relations to other variables, consequences of testing
Standards for educational and psychological testing (AERA et al., 2014)	*Sources of validity evidence*: test content, response processes, internal structure, relations to other variables, consequences of testing

> quality which is not 'operationally defined.' The problem faced by the investigator is 'What constructs account for variance in test performance?' (p. 282)

There is a bit of tautology in these definitions in that individuals' responses to test items reflect the construct the test is thought to measure. But clearly a notion of the construct must come before the creation of the test.

Although they emphasized the importance of construct validity, it is important to note Cronbach and Meehl (1955) did *not* assert that construct validity applied to all types of tests. Instead, they introduced the concept to address concerns regarding how the more psychological tests, such as projective and personality tests, could be validated. Certainly it would be difficult to evaluate such tests from content or criterion-related perspectives. As Cronbach and Meehl described, "Construct validity was introduced in order to specify types of research required in developing tests for which the conventional views on validation are inappropriate" (p. 299). The notion of "something" being measured, such as a latent trait or a construct, is needed in non-educational testing environments where specific curricular goals and content specifications do not exist. For this reason, Cronbach and Meehl asserted that construct validity "...must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured" (p. 282).

Loevinger (1957) argued that constructs are always implied when people are tested and there is no operational definition, criterion, or universe of content sufficiently adequate to remove the need for a construct interpretation of test scores. In describing construct validity in this way, Loevinger cemented the notion that all validity was construct validity. She claimed, "since predictive, concurrent, and content validities are all essentially *ad hoc* construct validity is the whole of validity from a scientific point of view" (p. 636). This perspective gained momentum, and gradually became the accepted consensus (cf. Ebel, 1961). After all, what psychometrician wants to be accused of being non-scientific? However, this consensus took a while. For example, the next version of the *Recommendations*, entitled *Standards for Educational and Psychological Tests and Manuals* (AERA et al., 1966), took a small step toward the idea that validity was a unitary concept by changing the four "types" of validity named in the 1954 *Recommendations* (content, concurrent, predictive, and construct) to three "aspects" of validity (content, criterion-related, and construct). The change from "types" to "aspects" was subtle, but clearly acknowledged the growing notion that there were not separate, but equal types of validity.

The next version of the *Standards* — the *Standards for Educational and Psychological Testing* (AERA et al.,

1974), took a major step forward in promoting the idea that validity was a unitary concept. They stated,

> The kinds of validity depend upon the kinds of inferences one might wish to draw from test scores...interdependent kinds of inferential interpretation are traditionally described to summarize most test use: the *criterion-related validities...content validity*; and *construct validity*...These aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation. A thorough study of a test may often involve information about all types of validity (AERA et al., 1974, pp. 25-26).

These Standards promoted a unitary view of validity, but retained the different aspects involved in validation (Cronbach, 1971). However, they did not go so far to say the unitary conceptualization was centered on construct validity. The next version of the *Standards*, now called the *Standards for Educational and Psychological Testing* (AERA et al., 1985) more explicitly promoted the idea of a unitary concept, but stopped short of claiming construct validity was the unifying force, as illustrated in the following excerpt,

> Validity...is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself (AERA et al., 1985, p. 9).

Thus, by the mid-1980s, the consensus was set that validity was a unitary concept. Around this same time, a particularly compelling and influential validity theorist, Samuel Messick, drew from Cronbach and Meehl (1955), Loevinger (1957), and others (e.g., Guion, 1977) to argue that, in essence, this unitary conceptualization was construct validity (Messick, 1975, 1980, 1988, 1989). In his landmark chapter in the third edition of *Educational Measurement* (Messick, 1989), he used philosophy, logical argument, and a comprehensive review of the literature and practice in educational and psychological testing, to claim that all interpretations of test scores, and the evaluation of the use of a test, must be viewed in relation to the construct the test intends to measure.

Messick's view of construct validity as the unifying force is similar to Loevinger's (1957) view; however, in discussing Loevinger's quote about criterion and content validity being "ad hoc," Messick (1989) commented,

> This chapter goes further still...it is here maintained that...reliance on criterion validity or content coverage is not enough. The meaning of the measure, and hence its construct validity, must always be pursued — not only to support test interpretation, but also to justify test use (p. 17).

To Messick, the process of testing implied measurement of a construct and so all validity was inherently construct validity. As he put it, "...the essence of the unified view of validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force is empirically grounded construct interpretation" (Messick, 1989, p. 64). Clearly, his intent was to eliminate the notion of separate types of validity once and for all. And to many validity theorists (e.g., Guion, 1980), he succeeded.

2 Beyond Construct Validity Theory: Test Validation

Messick's (1989) chapter on validity is perhaps the most comprehensive treatment of the topic ever written. The chapter is organized using two facets — the source/justification of testing, and the function/outcome of testing. Validity is described with respect to six philosophical orientations (logical positivism, relativism, rationalism, instrumentalism, realism, constructivism) and five systems of inquiry (Leibniz, Locke, Kant, Hegel, and Singer). Through all of these "philosophical conceits" (p. 21), Messick defends the unitary conceptualization of validity as construct validity. As he put it, "if construct validity is considered to be dependent on a singular philosophical base such as logical positivism and that basis is seen to be deficient or faulty, then construct validity might be dismissed out of hand as being fundamentally flawed" (p. 22).

Although Messick's scholarly arguments were compelling, there were some detractors who believed his formulation was either too restrictive (Ebel, 1977; Sireci, 1998; Yalow & Popham, 1983), or too obtuse to promote adequate test validation practices (Shepard, 1993). These criticisms most likely stemmed from his staunch defense of the idea that all validity is construct validity. For example, although Messick (1989) concluded all validity is construct validity, he also stated "validity is a unitary though faceted

concept" (p. 14), and that the distinctions among the facets are "not only interlinked but overlapping" (p. 20). Terms like "unified faceted" and "interlinked but overlapping" appropriately describe Messick's conceptualizations, but they are conceptually complex, and they do not inspire practitioners and lay audiences to learn more about validity. For this reason, Messick' s (1989) chapter has been criticized as inaccessible and lacking in guidance for applied validation purposes.

For example, in critiquing Messick's chapter, Shepard (1993), complimented him on his major points, but asked for a "simpler model for prioritizing validity questions, one that clarifies which validity questions must be answered to defend a test use and which are academic refinements that go beyond the immediate, urgent questions" (p. 407). Shepard agreed with Messick's theoretical points regarding construct validity, but thought it was not helpful for guiding validation practice. Instead, she encouraged test evaluators to "...ask directly 'What does a testing practice claim to do?' and to organize the gathering of evidence around this question" p. 408). This approach is compelling to both measurement practitioners and lay audiences. Although Shepard's chapter supported Messick's theory of construct validity, it is important for us to note that her suggestion to begin the validation inquiry by evaluating the claims of testing does not require use of the esoteric nomenclature of construct validity theory.

Another "de-constructed" approach to test validation is the argument-based approach promulgated by Kane (1992, 2006, 2013). Kane (1992) sidestepped a theoretical formulation of validity by providing a process for demonstrating a sufficient body of evidence to support test use and interpretation. Borrowing from Cronbach (1971, 1988), he proposed identifying the intended uses and interpretations of test scores and utilizing them as the framework for validation. His "argument-based approach" involved the establishment of an "interpretive argument as the framework for collecting and presenting validity evidence" (p. 527). The interpretive argument uses the "evidence categories" of Observation, Generalization, Extrapolation, and Theory to develop an argument to support the use or interpretation of test scores for specific purposes. This evidence-based argument is termed a "validity argument," which represents an evaluation of the reasonableness of the interpretive argument.

Kane (2006) defined the argument-based approach to validation as a two-step process. First, the proposed interpretations and uses of test scores are made explicit through the interpretive argument. Next, the elements of that interpretive argument are evaluated to develop the validity argument. This approach provides a practical means for evaluating the use of a test for a particular purpose, without invoking different types or aspects of validity. Kane (2013) extended his argument-based approach to include a "interpretation and use argument," and described his logic in developing the approach as,

> The argument-based approach was intended to avoid the need for a fully developed, formal theory required by the strong program of construct validity, and at the same time to avoid the open-endedness and ambiguity of the weak form of construct validity in which any data on any relationship involving the attribute being assessed can be considered grist for the mill (pp. 8-9).

By "strong" versus "weak" programs of construct validity Kane was referring to Cronbach's (1989) distinction that an ideal (strong) approach to construct validation would be driven by formal theories and series of hypothesis tests that could be continual; whereas an insufficient (weak) approach would focus on easily available data and "a miscellaneous collection of marginally relevant findings" (Kane, 2013, p. 7). Thus, the argument-based approach to validation acknowledges the fact that an ideal validation effort is typically not possible, but nevertheless a sufficient body of evidence must be put forward to support the use of a test for a particular purpose.

Kane's argument-based approach was essentially endorsed by the current version of the *Standards for Educational and Psychological Testing* (AERA et al., 2014), which stated, "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). With respect to validation, the *Standards*' adoption of the argument-based approach is clear. For example, they state:

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses...a test interpretation for a given use rests on evidence for a set of propositions making up the validity argument, and at some point validation evidence allows for a summary judgment of the intended interpretation that is well supported and defensible (pp. 21-22).

Although the AERA et al. (2014) *Standards* promote the idea of using a validity argument, they do not use the

same language proposed by Kane (1992, 2006, 2013). That is, they do not require the development of an interpretive argument. They also do not describe validity within a construct-based perspective. However they do require a clear definition of the construct measured by a test, and explicit statements of the testing purposes. As is implied in the *Standards*' definition of validity, validation is described as the process of providing evidence to support explicitly stated testing purposes. Thus, the argument-based approach as characterized in the AERA et al. (2014) the *Standards* is inherently practical.

To provide guidance for developing a validity argument, that is, for defending the use of a test for a particular purpose, the AERA et al. (2014) *Standards* stipulated five sources of validity evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). The sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. Note that in establishing this validation framework, the Standards focused validation on test interpretation and use, and avoided different types or aspects of validity — including construct validity. In fact, like AERA et al. (1985), the past two versions of the *Standards* defined validity as a unitary concept without granting authoritative status to construct validity. For example, in describing the five sources of evidence, they stated,

> These sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use (pp. 13-14).

2.1 Mislevy's Sociocognitive Perspective

Before leaving our historical tour of the validity literature, it is important to acknowledge the work of Mislevy (e.g., 2009, 2018), who expanded Messick's (1989) social considerations in test use by framing validation within a "sociocognitive" perspective, and discussed how differences in the psychometric modeling of an assessment imply and require different formal connections between the model and the inferences derived from test scores. As Mislevy (2009) described, "An essential element of test validity is whether, in a given application, using a given model provides a sound basis for organizing observations and guiding actions in the situations for which it is intended" (p. 83). This perspective on validity emphasizes test use, as do the AERA et al. (1999, 2014) *Standards*; and like Messick, it also acknowledges the variations in the social context in which assessment occurs.

To borrow terms from the philosophy of science, Mislevy's socicognitive perspective supports "...a constructivist-realist view of validity" (Mislevy, 2009, p. 84), which was also discussed by Messick (1989). The perspective is "realist" in that it presumes what is being measured really exists, but "constructivist" in that it acknowledges the conceptualization and measurement of the construct being measured can vary widely across test developers, measurement conditions, and context. As Mislevy (2009) described, "The constructivist-realist view holds that models are human constructions, but successful ones discern and embody patterns that characterize aspects of more complex real-world phenomena" (p. 95). Thus, the sociocognitive perspective is congruent with the *Standards*' definition that validity refers to the degree to which use of a test for a particular purpose is justified by evidence and theory, and thus it is more useful than narrower conceptualizations that are from a purely realist perspective (e.g., Borsboom et al., 2004).

2.2 Summary of the Evolution of Validity Theory and Standards

Our brief journey through the history of construct validity theory and the evolution of the *Standards for Educational and Psychological Testing* leaves us at a very practical place. Because validity refers to the degree to which evidence and theory support the use of a test for a particular purpose, validation then involves gathering and evaluating evidence focused on justification of the use of test scores for that purpose. Thus, the AERA et al. (2014) *Standards* provide us with a framework for gathering and organizing validity evidence. They remind us that validity must be evaluated with respect to specific uses of test scores, and emphasizes that one type of evidence is not likely to constitute a sufficient validity argument. It does all this without *relying* on the notion of a construct. Now that this de-constructed approach to test validation has been legitimized, I illustrate how it can be applied in practice.

3 Using the AERA et al. (2014) *Standards* as a Validation Framework

In Sireci (2013), I proposed a three-step validation process that used the AERA et al. (2014) *Standards* as a framework for validation. These steps involved: (a) clear articulation of testing purposes, (b) considerations of potential test

misuse, and (c) crossing test purposes and potential misuses with the *Standards*' five sources of validity evidence. However, I have added a fourth step to acknowledge the fact that it is not possible to conduct all validity studies that could theoretically be conducted. This fourth step is, (d) prioritizing the validity studies to be conducted. This last step is needed to ensure the validity argument is founded on evidence that focuses on whether use of the test for its intended purposes leads to more positive than negative outcomes; that is, evidence that use of the test does more good than harm. Next, I describe this four-step process.

3.1 Step 1: Articulating the Purposes of the Test

Validation involves gathering and analyzing evidence that bears on the defensibility of use of a test *for a particular purpose*. Thus, validation starts with identifying how test scores are used. As the AERA et al. (2014) *Standards* describe, "Validation logically begins with an explicit statement of the proposed interpretations of test scores, along with a rationale for the relevance of the interpretation to the proposed use" (p. 11). In most contexts, this initial step simply restates the purpose statements from a particular testing program. In fact the *Standards* require testing agencies to clearly specify the intended purposes of a test. As they put it, "The test developer should put forth clearly how test scores are intended to be interpreted and consequently used" (p. 23). The *Standards* also require test developers to clearly describe the construct measured with respect to the testing purpose (p. 85). Defining the construct measured and the testing purpose sets the foundation for validation.

Identifying the intended purposes of a test should be clear from its technical documentation (AERA et al., 2014, p. 125). However, in many cases the intended purposes of a test are complex or unclear. In such cases, the intended purposes must be derived from the explicit claims made by a testing agency. Once the intended purposes of the test are clearly articulated, which includes understanding the intended uses of test scores, validity studies to evaluate those uses can be proposed.

3.2 Step 2: Identifying Potential Negative Consequences of Test Use

Although it is important to focus validation on the intended uses of test scores, consideration of potential negative effects of the testing program are also critical (Messick, 1989). One way to identify potential negative effects is to follow public criticisms of a testing program. For example, achievement tests that are used as a high school graduation requirement are often criticized as "narrowing the curriculum" and stressing out students so much that they may drop out of high school. "Adverse impact," where the percentages of examinees passing the test differ by demographic factors such as sex and ethnicity, is another criticism and so a potential negative consequence to be evaluated. Potential negative consequences represent hypotheses to be studied — and those studies should be included in the validation framework.

3.3 Step 3: Crossing Test purposes and Potential misuses with the *Standards*' Five Sources of Validity Evidence

This step involves explicit inclusion of the AERA et al. (2014) *Standards*' five sources of validity evidence — validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations with other variables, and (e) testing consequences. Full description of each source of evidence is beyond the scope of this article and so readers are encouraged to refer to the Standards for more complete descriptions. Here, I present only very brief descriptions.

Validity evidence based on test content refers to studies that evaluate the degree to which the content of a test adequately represents the content tested and is consistent with the testing purpose (Martone & Sireci, 2009; Sireci & Faulkner-Bond, 2014). Such evidence is typically gathered using subject matter experts who review and rate test items with respect to the targeted content domain, or who help define that domain (e.g., job analyses used in licensure testing).

Validity evidence based on response processes refers to "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers" (AERA et al., 2014, p. 15). Such evidence can include interviewing test takers about their responses to test questions, systematic observations of test response behavior, evaluation of the criteria used by judges when scoring performance tasks, analysis of item response time data, and evaluation of the reasoning processes examinees use when solving test items (Embretson, 1983; Messick, 1989; Mislevy, 2009).

Validity evidence based on internal structure refers to statistical analysis of item and sub-score data to evaluate the degree to which the dimensionality of assessment data is congruent with the hypothesized dimensionality specified by the theory underlying the test and the statistical model used to score the test (e.g., a unidimensional item

response theory model). Statistical procedures like factor analysis (both exploratory and confirmatory), multidimensional scaling, or model-based residual analyses can be used to evaluate whether the hypothesized dimensionality is represented in examinees' responses to test items. The degree to which sub-scores are supported by dimensional analysis is also relevant here, as are studies of differential item functioning.

Validity evidence based on relations to other variables refers to traditional forms of criterion-related validity evidence such as concurrent and predictive validity studies, as well as more comprehensive investigations of the relationships among test scores and other variables such as multitrait-multimethod studies (Campbell & Fiske, 1959) and score differences across different groups of students, such as those who have taken different courses. These external variables can be used to evaluate hypothesized relationships between test scores and other measures of student achievement (e.g., test scores and teacher grades), to evaluate the degree to which different tests actually measure different skills, and the utility of test scores for predicting specific criteria.

Finally, *validity evidence based on consequences of testing* refers to studies of the intended and unintended consequences associated with a testing program. Examples of these types of were described earlier in Step 2, but further examples can be found in Messick (1989), Shepard (1993), and Lane (2014). In a sense, all validity studies can be thought of as evaluating the consequences of testing because testing purposes represent intended consequences. However, this category of evidence typically focuses on evaluating whether there are unintended negative effects associated with a test.

An example of Step 3, crossing the testing purposes and potential misuses with the AERA et al. (2014) *Standards*' five sources of validity evidence, is presented in Table 2. This example comes from the Massachusetts Adult Proficiency Tests (MAPT), which are math and reading tests for adult education students in Massachusetts. The purposes of these tests are explicitly stated in its *Technical Manual*, as "The purposes of the MAPT are to measure [adult education students'] knowledge and skills in mathematics and reading so that their progress in meeting educational goals can be evaluated...[it] is designed to measure learners' educational gains for the purposes of state monitoring and accountability" (Zenisky et al., 2018, p. 10). This purpose statement immediately suggests several types of validity questions that should be investigated to support use of the MAPT for these purposes. These questions are,

1. Does the MAPT actually measure adult education students' knowledge and skills in math and reading?
2. Does it measure these knowledge and skills as they are defined in the Massachusetts adult education curriculum frameworks?
3. Do MAPT scores provide accurate information regarding students' math and reading proficiencies?
4. Are MAPT scores useful for evaluating students' progress toward meeting educational goals?
5. Are MAPT scores appropriate for evaluating student progress as defined by the Federal Government?
6. Are aggregated MAPT scores useful for evaluating the effectiveness of ABE programs?

In addition the MAPT *Technical Manual* (Zenisky et al., 2018) identifies two potentially problematic uses and warns against them: use of the test for diagnostic purposes and for placing students into instructional programs. These two test uses suggest additional validity questions bearing on potential negative consequences. Specifically,

7. Are teachers inappropriately using MAPT scores to diagnose student's strengths and weaknesses?
8. Are adult education programs using the MAPT for placement purposes?
9. What are the effects of the MAPT on instruction in adult education?

The last question does not stem from any of the explicit testing purposes or warnings against inappropriate use. Rather, it emanates from an implied, more altruistic, purpose, that educational tests should be integrated with and improve instruction. These 9 validity questions form the rows in Table 2, and the five sources of validity evidence form the columns. These rows and column provide a framework that links the *Standards*' five sources of validity evidence to validity questions that automatically arise from the stated testing purposes and warnings. The check marks (√) in the table indicate where evidence is needed to address each validity question. The specific studies represented by the check marks in each cell are not described here (see Sireci, 2012 for examples), but many can be readily inferred. For example alignment studies would exemplify studies associated with the first cell in the table.

3.4 Step 4: Prioritizing the Validity Studies to be Conducted

The cells in Table 2 represent a set of studies that represent a comprehensive test validation agenda. However,

Table 2. Illustration of Crossing Test Purposes and Potential Misuses with the *Standards*' Five Sources of Validity Evidence

Validity Question	Source of Validity Evidence				
	Content	Internal Structure	Relations to External Variables	Response Processes	Testing Consequences
Measure correct skills?*	√ *	√ *	√ *	√	
Congruent with frameworks?	√ *				
Accurate?		√ *	√		
Measure progress?	√ *	√ *	√ *		
Meet Federal requirements?	√ *	√			
Useful for program Evaluation?	√	√			√
Inappropriate diagnostic use?					√
Inappropriate placement?					√
Effect on instruction?	√				√

*prioritized studies.

due to limited time and resources, it is typically not possible to implement such a comprehensive set of studies, and so some prioritization is needed.

It is important to note that all of the validity questions gleaned from the MAPT purpose statements are addressed in some form in Table 2, and all involve at least one sources of validity evidence. The prioritization of the validity questions must consider the most important purposes of the MAPT, and the primary reason it was created to meet the Federal accountability regulations. According to the Federal regulations, Massachusetts must have an assessment linked to its curriculum framework, and *must* use the assessment to evaluate students' educational gains according to the NRS achievement levels. Such assessment *must* be accurate, and is, by definition, used for evaluating programs. Thus, the validity studies to be prioritized must represent the minimum amount of evidence needed to argue that the MAPT is fulfilling its intended purposes. The prioritization of these studies to support a sufficient validity argument is denoted by the asterisks (*) in Table 2. As is evident from the asterisks, studies focused on ensuring the content is appropriate and Federal demands are met are prioritized.

3.5 Summary of Four-Step Validation Process

The proposed four-step process to guide the validation process, like the argument-based approach to validity, represents a compromise between carrying out all the validity studies that could theoretically be done to evaluate the use of a test for a particular purpose, and the minimum required to justify the use of a test for a particular purpose. There are some limitations of this approach. For one, it requires responsible test developers and evaluators to clearly articulate testing purposes and intended uses, identify potential misuses, and conduct quality studies that will provide the intended information. Second, it requires prioritization of studies, and it may be difficult to get agreement on this prioritization. However, a benefit of the process is it provides a standardized way for approaching validation that is grounded in the AERA et al. (2014) *Standards*, which are based on almost seven decades of scholarly collaboration. It also focuses on use of the test and the effect of such use. Thus, it is consistent with the AERA et al. definition of validity, and its advice regarding validation as an endeavor to justify the use of a test for a particular purpose.

4 Discussion

For over 50 years, the notion of a construct has been central to educational and psychological measurement. Thus, it is not surprising that many test specialists accepted construct validity theory as the most philosophically correct description of validity. In this article, I argued that we could use many of the tenets of construct validity theory to design and conduct validation efforts, without getting bogged down in the nomenclature. By "de-constructing" validation, the work becomes less philosophical and more

practical. Rather than focus on a construct, we focus on testing purposes and specific use of test scores. One way to incorporate this focus is the four-step validation process outlined in this article.

This approach to validation starts with clearly articulating the purposes of testing. Such articulation lays the foundation for the validation effort by implying the critical validity questions to be addressed. These questions should be subsequently prioritized to establish a validation agenda and timetable. This prioritization should also consider criticisms raised against the testing program, and the constraints in gathering validity evidence. Some constraints may be due to financial and personnel resources, while others may be due to the need to wait a period of time before sufficient test score data are available, or until the effects of testing have had time to occur. These constraints should not be used as excuses to halt validation efforts prematurely. Rather, they should be part of the discussion of how to answer the important validity questions in a sufficient and systematic way, given the resources at hand. Finally, the validation plan should be periodically updated as the testing program matures and as insights from previously conducted studies point to important new directions of inquiry.

In establishing the validation plan, we can draw helpful advice from the validity literature. One of my favorite quotes regarding validity, was supplied, of course, by Messick (1989) who stated,

> Tests are imperfect measures of constructs because they either leave out something that should be included according to the construct theory or else include something that should be left out, or both (p. 34).

This quote is important for validation plans because if validation endeavors strive to ensure tests have not left anything out (e.g., have adequately represented the content domain), and do not contain any sources of bias, we will have gone a long way in supporting the use of a test for a particular purpose.

A question remains regarding when a program of validity research has gathered a sufficient amount of evidence to support the use of a test for a particular purpose. In considering how much validation to do before we can be satisfied test scores are appropriately fulfilling their purposes, I have earlier suggested using a courtroom analogy. In Sireci (2009), I put forward the following suggestion:

> Perhaps the best question to guide test validation efforts is "If the use of this test for the purpose I am using it for were challenged in court, do I have sufficient evidence to persuade the judge or jury and win the case?" If the answer is yes, the evidence will comprise a solid validity argument...If not, more evidence is needed, or use of the test cannot be defended (p. 31).

5 Conclusions

Clearly, evidence that the test is measuring the intended construct is important in validation. However, that evidence can be put forward and organized by at least four of the five sources of validity evidence promulgated by the AERA et al. (2014) *Standards*. In this article I argued we can conduct proper and sufficient validity investigations without using the term "construct validity." Construct validity theory is elegant from a philosophy of science perspective, but even Messick (1989) conceded, "the philosophy of science is more philosophy than science" (p. 21). Hopefully, the practical guidance outline in the four-step validation process presented in this article will lead to the development of sufficient validity arguments that support the use of tests for specific purposes, and lead to testing programs that benefit society with minimal negative consequences.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

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

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

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2, Pt.2), 1–38.

American Psychological Association. Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist*, *7*, 461–475.

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

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

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, *22*(3), 5–11.

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

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

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

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

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist*, *16*, 640–647.

Ebel, R. L. (1977). Comments on some problems of employment testing. *Personnel Psychology*, *30*, 55–63.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Guion, R. M. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, *1*, 1–10.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, *11*, 385–398.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*, 127–135. doi: 10.7334/psicothema2013.258

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694 (Monograph Supplement 9).

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research 4*, 1332–1361.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, New Jersey: Lawrence Erlbaum.

Messick, S. (1989). Valifity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Washington, D.C.: American Council on Education.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83–108). Charlotte, NC: Information Age Publishing Inc.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*, 301–319.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405–450.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83–117.

Sireci, S. G. (2009). Packing and unpacking sources of

validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing Inc.

Sireci, S. G. (2012). *Smarter balanced assessment consortium: Comprehensive research agenda.* Available at http://www.smarterbalanced.org/assessments/development/additional-technical-documentation/.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*, 99–104.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*, 100–107. doi: 10.7334/psicothema2013.256

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York, NY: Teachers College Press.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher*, *12*, 10–14.

Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O' Donnell, F., Wells, C. S., ... Garcia, A. (2018). Massachusetts adult proficiency tests for college and career readiness: Technical manual. *Center for Educational Assessment research report No. 974*. Amherst, MA: Center for Educational Assessment.