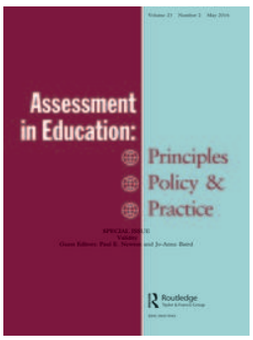

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

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

On the validity of useless tests

Stephen G. Sireci

To cite this article: Stephen G. Sireci (2016) On the validity of useless tests, Assessment in Education: Principles, Policy & Practice, 23:2, 226-235, DOI: 10.1080/0969594X.2015.1072084

To link to this article: http://dx.doi.org/10.1080/0969594X.2015.1072084

Assessment in Education: Principles, Policy & Practice, 2016
Vol. 23, No. 2, 226–235, http://dx.doi.org/10.1080/0969594X.2015.1072084

On the validity of useless tests

Stephen G. Sireci*

College of Education, University of Massachusetts Amherst, Amherst, MA, USA

(*Received 7 January 2015; accepted 9 July 2015*)

A misconception exists that validity may refer only to the *interpretation* of test scores and not to the *uses* of those scores. The development and evolution of validity theory illustrate test score interpretation was a primary focus in the earliest days of modern testing, and that validating interpretations derived from test scores remains essential today. However, test scores are not interpreted and then ignored; rather, their interpretations lead to actions. Thus, a modern definition of validity needs to describe the validation of test score interpretations as a necessary, but insufficient, step en route to validating the *uses* of test scores for their intended purposes. To ignore test use in defining validity is tantamount to defining validity for 'useless' tests. The current definition of validity stipulated in the 2014 version of the *Standards for Educational and Psychological Testing* properly describes validity in terms of both interpretations and uses, and provides a sufficient starting point for validation.

Keywords: assessment; educational testing; psychological testing; testing standards; validity

As the articles in this special issue illustrate, there have long been debates about the meaning of *validity* as it applies to educational and psychological testing. One issue of debate is whether validity refers to the *interpretations* of test scores or to the *use* of test scores. In reviewing the history of this debate, two points stand out that help explain the disagreements and suggest how to rectify them. The first is the lack of attention paid to the important distinction between *necessary* components of validity and a *sufficient* validity argument. The second is the misconception that test scores can be interpreted without ever being used for a particular purpose. In the remainder of this article, I elaborate on these two points to explain the sources of disagreement over what *validity* is and how we can get beyond this debate to move forward as a united community with common understandings of test validity and validation.

The earliest definitions of validity

There are several comprehensive accounts of validity history (e.g. Kane, 2013; Messick, 1989; Newton & Shaw, 2013; Sireci, 1998; Sireci & Sukin, 2013) and so I will not provide an exhaustive review here. However, I review a bit of this history to illustrate how different definitions of validity emerged and how the distinction

*Email: sireci@acad.umass.edu
Paper to appear in special issue of *Assessment in Education: Principles, Policy & Practice*, J. Baird and P. Newton, editors.

between necessary and sufficient components of validity was underemphasised, which led to disagreements over its meaning. I begin in the early twentieth century.

As educational and psychological tests emerged at the beginning of the twentieth century, there were two early definitions of validity. One described validity as the degree to which a test 'measures what it purports to measure' (e.g. Garrett, 1937; Smith & Wright, 1928); the other referred to validity in correlational terms, such as Guilford (1946) who succinctly stated, 'a test is valid for anything with which it correlates' (p. 429; see also Bingham, 1937; Kelley, 1927; Thurstone, 1932, for similar definitions). The first definition led to validity studies based on factor analysis to determine whether the underlying 'factors' discovered through analysis of examinees' responses to test items conformed to the hypothesised conceptualisation of the 'construct' measured. The second definition led to criterion-related validity studies to determine whether test scores were consistent with other measures of the construct (knowledge, skill or psychological trait) tested.

These two definitions were not mutually exclusive. For example, Newton and Shaw (2013) pointed out that, in 1921, the National Association of Directors of Educational Research reported the results of a survey of its members that sought to standardise emerging terms in the educational research field. Newton and Shaw provided an excerpt from the *Standardisation Committee*'s report that illustrates the definition of validity in terms of what a test measures. The excerpt they cited was,

> Two of the most important types of problem in measurement are those connected with the determination of what a test measures, and of how consistently it measures. The first should be called the problem of validity, the second, the problem of reliability. (Buckingham et al., 1921, p. 80; cited in Newton & Shaw, 2013)

However, the next sentence from that same report indicates the Committee also considered validity to be determined through relations of test scores with other measures of the same construct. That sentence reads,

> Members are urged to devise and publish means of determining the relation between the scores made in a test and other measures of the same ability; in other words, to try to solve the problem of determining the validity of the test. (Buckingham et al., 1921, p. 80)

Defining validity in terms of 'what the test measures' was important to psychometricians in the early twentieth century because they were not only concerned with 'validating' a particular test, they were also concerned with justifying the practice of psychological measurement. That is, they needed to justify the new practice or 'science' of psychological measurement by demonstrating 'what' they were measuring actually existed. Thus, demonstrating that the interpretations derived from these new assessments reflected 'real' attributes of the individuals who took them was a major focus of 'validation.' The statistical methods available at that time – correlation and factor analysis, were extensively used to confirm a test measured 'what' it was supposed to measure.

These two definitions began the academic discussion of what validity was and how tests should be validated. As part of that discussion, it was noted that efforts to validate tests based on statistical analyses were incomplete because the composition of the test (e.g. its content and other qualitative attributes) was ignored. This concern led to an extension of validity to include an appraisal of how well the content of a test represented the intended construct and testing purpose (e.g. Rulon, 1946). Thus, by the 1940s, definitions of validity included (a) appraising the degree to which a

test measured what it claimed to measure, (b) evaluating the degree to which test scores correlated with other measures of the intended construct, (c) evaluating the consistency of test content with the goals of testing and (d) evaluating how well the test scores were useful for specific purposes.

The idea that validity referred to more than statistical analysis of test or item scores was an important early development in the evolution of validity theory because it emphasised a broader conceptualisation that focused on how test scores were used and their defensibility for specific uses. Pressey (1920) was one of the first to point out the limitations of a purely statistical approach to validation. As he described,

> Our statistical methods as applied to tests have been largely borrowed … from the descriptive sciences. So the question has been: What is the test measuring, and how accurately is the thing being measured? But mental testing is not a descriptive, but a technical science. And the question should be instead: What are we trying to do and how well are we doing it? (p. 472)

Pressey's (1920) points were that 'what the test is measuring' is an insufficient validation question, and that validation required a broader conceptualisation of validity focusing on the intended use of the test. Rulon (1946) made this latter point explicit by stating 'we cannot label a test valid or not valid except for some purpose' (p. 290). These notions are consistent with contemporary definitions of validity as embodied in professional standards for educational and psychological testing. However, before moving to contemporary validity theory, it is important to note that the focusing of validation on the use of a test for a specific purpose, which began as early as 1920, does not negate the importance of ensuring tests are measuring what they intend to measure or demonstrating test scores correlate with other measures of the targeted construct. These earlier definitions persevere as important and necessary aspects of validity. They are expanded, not replaced, by modern validity theory, which acknowledges test scores are used for specific purposes, and it is these uses that need to be validated, not the test itself.

Efforts to establish an authoritative definition of validity

The first attempt to provide a consensus or authoritative definition of validity was the 'Technical recommendations for psychological tests and diagnostic techniques' (American Psychological Association [APA], 1954), which represented a collaborative effort among the three most influential educational and psychological professional associations in the United States: APA, the American Educational Research Association (AERA) and what today is the National Council on Measurement in Education (NCME). There have been five revisions of this document, the latest being the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

The first version of these *Standards* specified four different 'types' of validity: content, predictive, concurrent and construct (APA, 1954, p. 13). Predictive and concurrent validities fit the traditional notion of a test being validated with respect to a criterion, and construct validity fits the definition of appraising how well a test measured what it claimed to measure. The second and third versions of the *Standards* renamed the 'types' to 'aspects' of validity, and combined predictive and concurrent validities into 'criterion-related validity' (APA, 1966; APA, AERA, & NCME, 1974).

The use of 'types' or 'aspects' to describe validity was a somewhat fragmented conceptualisation, and so it is understandable how disagreements over the definition of validity emerged. However, it is important to note that the fundamental notions that (a) validity was *not* an inherent property of a test and (b) validation must focus on the intended purposes of the test were explicit from the beginning. As noted in the first version of the *Standards*,

> It is not appropriate to call for a particular level of validity and reliability … It *is* appropriate to ask that the manual give the information necessary for the user to decide whether the accuracy, relevance, or standardization of the test makes it suitable for [its] purposes. (APA, 1954, p. 2)

and

> No manual should report that 'this test is valid' … The manual should report the validity of each type of inference for which a test is recommended. If validity of some recommended interpretation has not been tested, that fact should be made clear. (APA, 1954, p. 19)

The fourth version of the *Standards* (AERA, APA, & NCME, 1985) moved away from a fragmented conceptualisation of validity and described validity as a 'unitary concept' (p. 9) stating it referred to 'the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores' (p. 9). Although this description of validity focused on inferences, rather than intended purposes or uses, the 1985 *Standards* also stated 'The inferences regarding specific *uses* of a test are validated, not the test itself' (p. 9, emphasis added).

This latter quote from the 1985 version of the *Standards* provides a good example of a lack of clarity regarding whether validity refers to inferences regarding test scores (i.e. interpretations of test scores) or to the actions that are made on the basis of those interpretations (i.e. uses of test scores). This lack of clarity is where many are currently stuck, with some arguing validity refers to interpretations (e.g. Cizek, 2012) others arguing validity refers to uses (e.g. Sireci, 2013) and others arguing it refers to both (e.g. Kane, 2013).

In the subsequent versions of the *Standards* (AERA, APA, & NCME, 1999, 2014) this lack of clarity was resolved through an explicit definition of validity that made it clear test interpretation and test use are inseparable. As stated in the most recent version, 'Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (AERA et al., 2014, p. 11). This definition properly places interpretation as a temporary step en route to the more permanent actions associated with test use. It also acknowledges the reality that we cannot have test score interpretation without test score use. Before I elaborate on this point, I take one more historical detour to illustrate how the idea that 'validity refers to interpretation' gained traction in the latter part of the twentieth century.

Messick's 'interconnected facets' of validity

Messick (1989) wrote the Validity chapter in the third edition of the book *Educational Measurement* (Linn, 1989). Many psychometricians consider this chapter to be one of the most important and influential treatises on validity ever written. The first sentence of the chapter provided a comprehensive definition of validity:

> Validity is 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 and other modes of assessment. (Messick, 1989, p. 13)

The definition is comprehensive in that it specifies the need for both theory and empirical evidence in validation, and it specifies the need to support both interpretations of test scores (inferences) and how they are used (actions).

As Messick's (1989) definition pointed out, validity is integrative in at least two ways. First, it involves a synthesis of evidence to evaluate a test. Second, it integrates concerns about the appropriateness of the interpretations of test scores with concerns about the uses of the test scores. Messick's definition explicitly specifies 'inferences and actions,' which like the AERA et al. (2014) *Standards*' definition of validity, acknowledges test score interpretation is always followed by some action (i.e. test score use).

Like many others (e.g. Kane, 2006; Shepard, 1993), I applaud Messick's (1989) definition of validity. However, in further describing validity in that seminal chapter, Messick made his conceptualisation nebulous by stating 'validity is a unified though faceted concept' (p. 14). In elaborating on this 'unified' concept he attempted to distinguish 'two interconnected facets' (p. 14), which he described as,

> One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or outcome of the testing, being interpretation or use. (p. 20)

The postulation that validity was unitary, but involved interconnected facets that were 'not only interlinked but overlapping' (Messick, 1989, p. 20) elevated the complexity of Messick's validity theory. Unfortunately, it also drew attention away from the solid definition he proposed at the outset of the chapter. I believe some psychometricians have focused on the interpretation part of the 'outcome-of-testing' facet at the expense of ignoring the fact that it is only a fraction of a unitary concept. If we look at only one-half of one facet of Messick's theory, we can conclude validity can refer solely to test interpretation. If we look at only the definition of validity in the APA (1954) version of the *Standards*, and ignore the rest of what it says about validating test use, we can reach the same conclusion. But we cannot look at only a portion of Messick's theory or cherry pick quotes from the *Standards* to support the notion that test interpretation can be separated from test use in any meaningful way.

To summarise the brief historical tour, if we look at the influential writings in the validity literature, we can see where some theorists have gotten the notion that validity refers solely to test score interpretation (e.g. Cizek, 2012). However, that conclusion can only be based on a purely academic argument that could never be of value in reality. I elaborate on the absurdity of this argument next.

Test interpretation without test use

Earlier, I pointed out that providing evidence that a test is measuring what it purports to measure is a necessary component of a validation effort and so it is certainly germane to validity. Such evidence helps us evaluate, and validate, the interpretations made on the basis of test scores. In fact, all five sources of

validity evidence stipulated in the AERA et al. (2014) *Standards* (test content, response processes, internal structure, relations to other variables and consequences of testing) should illuminate score-based interpretations. Therefore, it should be clear I am *not* arguing against validating interpretations of test scores. I am also *not* arguing that test score interpretation is irrelevant to validity or validation. I am arguing test score interpretation is *part* of validation, and *partly* what validity refers to. Validating interpretations of test scores is a *necessary* component of any validation endeavour. However, it is not *sufficient* for defending the use of a test for a particular purpose.

To understand my point, it is helpful to ask the question 'Can we have test interpretation without test use?' That is, can someone interpret a test score, but never act upon that interpretation? I must admit it is possible, but why would we develop tests that we expect will never be used for a practical purpose? Theoretically, a physician could interpret an X-ray and not act upon the information in subsequent treatment of a patient. Similarly, an elementary school mathematics teacher can interpret a child's maths test score, and ignore that information when planning instruction for that child. It would be natural to ask the physician or the teacher why they administered these assessments if they were not going to use them. But perhaps a rhetorical question is better – Would you want this physician as your doctor, or this teacher for your child?

Although these questions and examples may seem silly, they essentially represent what would be the state of affairs if it were relevant to validate test score interpretations without validating the use of test scores. If tests existed only for their scores to be interpreted, but the scores were never used for any purpose, by definition, they would be *useless tests*. Therefore, we can conclude validity refers solely to test score interpretation for useless tests, but for tests that are actually used for some purpose, validity refers to the appropriateness, soundness and utility of the actions that are made based on the interpretations (i.e. their uses).

At this point, the current definition of validity provided by the AERA et al. (2014) *Standards* bears repeating:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. (p. 11)

I do not see the need to break this definition into separate components (i.e. one for interpretations, another for uses) or to reduce it solely to score interpretation. Validity is sufficiently defined by including both test score interpretation and test score use. It is not sufficiently defined when test score use is ignored. If score use is ignored, then a test could theoretically be 'validated,' but as Jenkins (1946) rhetorically asked, validated 'for what?' (p. 93). As another seminal researcher put it 70 years ago, 'we cannot label a test valid or not valid except for some purpose' (Rulon, 1946, p. 290).

The interpretation versus use issue may be helped by stepping back from the psychometric literature and simply consulting a dictionary for the definition of validity. The Merriam-Webster online dictionary (2015), provides a shorthand definition for 'valid' as 'fair or reasonable' and 'acceptable according to the law.' The full definition provides four meanings:

(1) having legal efficacy or force; *especially*: executed with the proper legal authority and formalities <a *valid* contract>

(2) a: well-grounded or justifiable: being at once relevant and meaningful <a *valid* theory>
b: logically correct <a *valid* argument><*valid* inference>
(3) appropriate to the end in view: effective <every craft has its own *valid* methods>
(4) *of a taxon*: conforming to accepted principles of sound biological classification

Although the first three are relevant to validity in educational and psychological measurement, the second definition is perhaps most pertinent. Part 'b' refers to logical correctness, which seems to apply directly to test score interpretations (e.g. is the inference correct?). Part 'a' uses the adjective 'justifiable,' which would seem to apply directly to test score use. 'Well-grounded' in 2a could encompass both the theoretical rationale regarding the construct measured (interpretation) as well as empirical evidence to justify test use. My interpretation of the dictionary definition of validity is that separate definitions can exist for a valid interpretation and a valid use. However, with respect to educational and psychological testing, I believe we cannot be satisfied with separate definitions. We need both, or at least we need what is contained in 2a: 'being at once relevant and meaningful.' I would add 'being at once relevant (for a purpose) and meaningful (with respect to the interpretation)'. Before leaving the Merriam-Webster definition, it is interesting to note the legal connotations of validity. A valid argument will be competitive in the courtroom. And an argument-based approach to validity is needed to support the use of test scores for a particular purpose (Kane, 2006, 2013). As I have written elsewhere (e.g. Sireci, 2013) the argument should not be prescriptive, but rather should address the specific test use and be organised by prioritising the most relevant sources of validity evidence described in the AERA et al. (2014) *Standards*.

Looking forward: defining validity for test validation

In the preceding sections, I described why I believe some psychometricians (e.g. Cizek, 2012) argue that validity can simply refer to test score interpretation, and why that argument is specious. It makes sense for us to think of the validity of an interpretation, but it does not make sense for us to think of interpretation devoid of application. Thus, it is not helpful to our profession, or to the science and practice of psychometrics, to restrict validity to the theoretical realm of test score interpretation. Actions and uses need to be validated, not interpretations that exist somewhere in the netherworld, never to be acted upon.

Defining validity based solely on whether a test reflects manifestations of a construct, and pretending those reflections (i.e. scores) will never influence further actions, is like looking at a traffic light, seeing what colour it is and driving straight through without considering the colour of the light. For the rest of my life, I plan on using the information communicated to me by the traffic light. And for the rest of my career, I plan on working towards improving the validity of tests that *will* be used. I remain uninterested in tests whose scores will be interpreted, but not used for anything. Useless tests have no utility and proposing a definition of validity for them is a fruitless endeavour.

I propose we move past the academic debate over what validity refers to and accept the AERA et al. (2014) definition. This definition correctly stipulates the

integration of test interpretation and use, and thus provides an appropriate foundation to guide validation. Although the *Standards* were developed by professional organisations in the United States, they are internationally respected, and as Zumbo (2014) concluded 'the *Standards* play a key role in the test and assessment community worldwide' (p. 32).

In addition to the definition of validity provided by the AERA et al. (2014) *Standards*, other validity theorists, such as Kane (1992, 2006, 2013) have stressed the importance of connecting test interpretation and test use. Kane (2013) suggested separate validity arguments for test interpretation and test use, but I believe a separate argument for test interpretation is only helpful as a preliminary step in building the validity argument for test use. I continue to believe validation should be focused on the intended and actual uses of test scores, and involves determining the validity evidence that is needed to support those uses. To support the use of a test for a particular purpose will require evidence that the test is measuring its intended construct, the scores are interpreted as appropriate manifestations of that construct, and the actions based on those interpretations are defensible.

Defining validity with respect to how test scores are used, an idea which dates back to the early twentieth century (e.g. Kelley, 1927; Pressey, 1920) and that is supported by the AERA et al. (2014) *Standards*, does not negate the importance of validating inferences derived from test scores. All five sources of validity evidence are relevant to test score use, and all, with the possible exception of evidence based on the consequences of testing, are also relevant to test interpretation.

Validating test score interpretations is a *necessary* component of validation, but it is not *sufficient* for supporting the use of a test for a particular purpose. Similarly, the different sources of validity evidence may be necessary for some applications, but it is unlikely any one source alone will be sufficient to validate the use of a test for a particular purpose.[1] Insufficiency does not mean there is a weakness in any one source of evidence, or that validating score inferences is of little worth. The insufficiency of any one source of validity evidence, or of the notion that validity refers to only test score interpretation, merely reflects the complexity of educational and psychological assessment. If we do our due diligence in providing a comprehensive validity argument based on sufficient evidence to support the use of a test for a particular purpose, we can have confidence that the actions based on these test scores are justifiable. This call to focus on test use in validation is not a twenty-first-century development. It echoes the claim made by Kelley almost 90 years ago:

> The establishment of the fact that a given test is valid for a specifically named purpose is at present one of the most, if not in fact the most, difficult of the problems confronting the test deviser. (Kelley, 1927, pp. 30–31)

My hope is we can move away from the academic issue of whether test use needs to be part of validation and concentrate on gathering evidence for the use of a test for a specific purpose. Validating test use may still be a difficult endeavour, but the AERA et al. *Standards* provide a helpful validation framework. I am sure Kelley would be proud of the numerous, comprehensive technical manuals that support contemporary educational and psychological tests by including comprehensive validity arguments that include various sources of validity evidence targeted towards appropriate test use.

Disclosure statement

No potential conflict of interest was reported by the author.

Note

1. It is also unlikely that all five sources will be needed to validate a specific test use (although certainly more than one source will be needed – the specific sources required depend on testing purpose, see Sireci, 2012, 2013).

Notes on contributors

Stephen G Sireci, PhD is a professor of Educational Policy, Research, and Administration and director, Center for Educational Assessment, University of Massachusetts, USA. His primary research interests are educational test development, evaluating educational tests, computer-based testing and cross-lingual assessment.

References

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

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

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, (2, supplement), 461–475.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.

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

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York, NY: Harper.

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

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

Garrett, H. E. (1937). *Statistics in psychology and education*. New York, NY: Longmans, Green.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427–439.

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology, 10*, 93–98.

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

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

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

Kelley, T. L. (1927). *Interpretation of educational measurement*. Yonkers-on-Hudson, NY: World Book.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

Merriam-Webster. (2015). *Validity* (online dictionary). Retrieved July 6, 2015, from http://www.merriam-webster.com/dictionary/validity

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

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

Pressey, S. L. (1920). Suggestions looking toward a fundamental revision of current statistical procedure, as applied to tests. *Psychological Review, 27*, 466–472.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16*, 290–296.

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. (2012, December). *Smarter balanced assessment consortium: Comprehensive validity agenda*. Retrieved July 6, 2015, from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/Smarter-Balanced-Research-Agenda_Recommendations-2012-12-31.pdf

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

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger (Editor-in-chief), *APA handbook of testing and assessment in psychology* (Vol. 1, pp. 61-84). Washington, DC: American Psychological Association.

Smith, H. L., & Wright, W. W. (1928). *Tests and measurements*. New York, NY: Silver, Burdett.

Thurstone, L. L. (1932). *The reliability and validity of tests*. Ann Arbor, MI: Edwards Brothers.

Zumbo, B. (2014). What role does, and should, the test standards play outside of the United States of America? *Educational Measurement: Issues and Practice, 33*, 31–33.