VM²ED
Validity Evidence for Measurement in Mathematics Education

An Introduction to Validity in Educational and Psychological Testing

Timothy D. Folger[1], Katie Burkett[2], Jonathan Bostic[1], Erin Krupa[2], and Brianna Bentley[2]

[1]School of Inclusive Teacher Education, Bowling Green State University

[2]Department of STEM Education, North Carolina State University

If you would like to reference this material, then please use the following citation:

Folger, T. D., Burkett, K., Bostic, J., Krupa, E., & Bentley, B. (2024). *An introduction to validity in educational and psychological testing.* Validity Evidence for Measurement in Mathematics Education. https://www.mathedmeasures.org

This material is based upon work supported by the National Science Foundation under Grant No. (DRL #1920619 & #1920621). Any opinions, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The purpose of this document is to describe how validity has been conceptualized for the Validity Evidence for Measurement in Mathematics Education (VM²ED) project. The VM²ED project was undertaken to systematically identify and compile measures used within mathematics education contexts (i.e. assessments, inventories, scales, observation protocols, etc.) and any associated evidence of validity related to each measure. The term *test* refers to any "evaluative device or procedure in which a systematic sample of a test taker's behavior in a specified domain is obtained and scored" (AERA et al., 2014, p. 224). Thus, different measurement instruments (e.g., assessments and inventories) are broadly categorized as a test (AERA et al., 2014).

The VM²ED project adheres to the *Standards for Educational and Psychological Testing* (*Standards*) as a framework for conceptualizing validity and validation. The three most influential social science organizations in the United States of America pertinent to the field of measurement are the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council of Measurement in Education (NCME) (Sireci, 2016). The *Standards* represent a consensus among these organizations on how validity is defined and evaluated when developing and evaluating tests. However, it is also important to acknowledge the continued disagreement among experts in the field of measurement in defining validity (e.g., Borsboom et al., 2004; Borsboom & Wijsen, 2016; Cronbach, 1988; Folger et al., 2023; Kane, 2013, Markus, 2016; Newton & Shaw, 2016; Shepard, 2016; Sireci, 2016). The VM²ED project conceptualizes validity through four propositions which are explored in subsequent sections:

A. Validity is an attribute of test-score interpretation for proposed uses of tests;
B. Validity is a unitary concept;

C. There are five sources of evidence used to evaluate the validity of an intended test-score interpretation and use;

D. Validity evidence reflects claims relative to the interpretation of test scores for proposed uses of tests.

Validity is an Attribute of Test-Score Interpretation(s) for Proposed Uses of Tests

The *Standards* broadly define validity as the "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). In other words, validity is an attribute of how test scores are interpreted and subsequently used (AERA et al., 2014; Kane, 2013; Messick, 1995; Shepard, 2016). To interpret test scores is to ascribe meaning to the scores; test-score use refers to actions taken or decisions made based on what the scores mean (Folger et al., 2023; Messick, 1995). To be clear, validity is not an attribute of a test. It is incorrect to refer to 'test validity' or to imply that 'a test is valid' (AERA et al., 2014). Messick (1995, p. 741) used Cronbach's (1971) ideas about validity:

> Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores… what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails.

Validity is a Unitary Concept

Prior to 1950, validity was evaluated through a variety of procedures and referred to by a variety of names (Anastasi, 1986). For example, content validity focused on whether a test collectively represented some larger domain (e.g., Cronbach & Meehl, 1955; Kane, 2013) and criterion validity was defined as the correlation between test scores and criterion scores (e.g., Cronbach & Meehl, 1955; Cureton, 1965; Kane, 2013; Thurstone, 1931). Criterion validity could further be described as concurrent validity or predictive validity based on when criterion scores

were collected (e.g., Cronbach & Meehl, 1955; Kane, 2013). Over time, Loevinger's (1957) argument that these validities were ad hoc and could be collectively referred to as construct validity became largely accepted.

The VM²ED project adheres to the *Standards* definition of validity as a unitary concept. Although describing distinct types of validity (e.g., content validity or predictive validity) is an outdated practice, distinct types of validity maintain relevance to the five sources of validity evidence.

Five Sources of Validity Evidence

The degree to which an interpretation and use is valid depends on the quantity and quality of supporting evidence (AERA et al., 2014; Kane, 2013; Pellegrino et al., 2016; Sireci & Padilla, 2014). An intended interpretation and use of test scores is not evaluated dichotomously as valid or not valid. The *Standards* enumerate five sources of evidence, described in Table 1. Each source is likely to illuminate different aspects of validity (AERA et al., 2014; Melhuish & Hicks, 2020). For example, evidence based on relations to other variables may empirically support assertions that test scores predict some future performance, such as students' future grade point average. There are a multitude of ways in which researchers collect particular sources of evidence. For specific examples related to each source of evidence, refer to the VM²ED *Evidence Types Guidebook.*

A Note on Reliability

In collecting the validity evidence associated with measures used in mathematics education contexts, the VM²ED team also collected evidence related to the reliability of the measures. It should be noted that reliability is neither a source of validity evidence nor is it validity evidence (AERA et al., 2014; Kane, 2013). *Reliability* (or *precision*) is "the degree to

which test scores for a group of test takers are consistent over repeated applications of a measurement procedure… the degree to which scores are free of random errors of measurement for a given group" (AERA et al., 2014, pp. 222-223). This notion of consistency (i.e., reliability) is a necessary condition for validity because "almost all test-score interpretations involve generalizations over some conditions of observation" (Kane, 2013, p. 3). Evidence of reliability can have implications on the validity of test-score interpretation and use, but evidence of reliability by itself does not support the validity of test-score interpretation and use (AERA et al., 2014; Kane, 2013). Put simply, validity and reliability are distinct aspects of test development and are evaluated separately, but both are relevant in assessing the appropriateness of a test.

Table 1

Description of the Five Sources of Validity Evidence (AERA et al., 2014)

Evidence Source	Description
Test Content	Test content includes the wording and format of test items or tasks. Test content evidence indicates that test items, or test content, align to the construct a test intends to measure.
Response Processes	Response processes describes the alignment between test takers' performance or behavior and the construct a test intends to measure.
Internal Structure	Internal structure may indicate the degree to which test items conform to the construct a test intends to measure. Such evidence may be collected through analysis of test dimensionality and item interrelationships.
Relations to Other Variables	Relations to other variables examine the degree to which test scores are, or are not, related to some ancillary variable.
Consequences of Testing	Consequences of testing present the intended and unintended consequences following the interpretation and use of test scores. Unintended consequences warrant close examination, and consequential evidence may anticipate and proactively address unintended consequences.

Validity Evidence Reflects Claims Relative to the Interpretation and Use of Test Scores

The *Standards* indicate that the type of validity evidence needed to support an intended interpretation and use of test scores "can be clarified by developing a set of propositions or claims that support the proposed interpretation for the particular purpose of testing" (AERA et al., 2014, p. 12). Validation broadly refers to the process through which the validity of a test-score interpretation and use is evaluated (AERA et al., 2014). There are typically two types of claims that are evaluated during validation. First, claims arising from the test-score interpretation. For instance, the claim that a student is or is not gifted in mathematics (Cizek, 2016; Kane, 2013). Second, claims inherent to the intended test-score interpretation and use (Folger et al., 2023). For example, an assessment developer's claim that a test is a unidimensional measure of mathematical problem solving

Current best practices in validation include the following: (a) Describe how test scores are to be interpreted and used, (b) Identify the claims relevant to the interpretation and use, and (c) Gather validity evidence related to those claims (AERA et al., 2014; Folger et al., 2023). However, it is also possible—and at times necessary—to develop claims from validity evidence that has already been collected (Folger et al., 2023). In other words, although validity evidence supports claims relevant to test-score interpretation and use, claims may be developed from existing validity evidence. For example, results of a factor analysis—which would reflect validity evidence based on internal structure—may support claims of test dimensionality, regardless of when the claim is formally presented.

Conclusion

Validity is a fundamental aspect of test development and test evaluation (AERA et al., 2014). Yet, validity and validation remain ambiguous across research communities (e.g., Folger

et al., 2023; Newton & Shaw, 2016; Wolming & Wikstrom, 2010), including the mathematics education research community (e.g., Bostic et al., 2021; Carney et al., 2022; Hill & Shih, 2009; Krupa et al., 2021). A goal of this paper is to define validity, and describe how validity has been conceptualized for the Validity Evidence for Measurement in Mathematics Education project. In doing so, this document provides scholars and practitioners with an introduction to validity and validation scholarship. Users of the VM²ED repository may leverage this information about validity and validation when identifying and selecting instrument(s) to use for a particular purpose of testing. Additionally, test developers and others engaged in validation work may consider leveraging the described conceptualization of validity during test development and test evaluation.

References

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

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

Anastasi, A. (1986). Evolving concepts of test validation. *Annual review of Psychology*, *37*(1), 1-16.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061.

Borsboom, D., & Wijsen, L. D. (2016). Frankenstein's validity monster: The value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 281-283.

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education, 24*(1), 5-31.

Carney, M. B., Bostic, J., Krupa, E., & Shih, J. (2022). Interpretation and use statements for instruments in mathematics education. *Journal for Research in Mathematics Education, 53*(4), 334-340.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212-225.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Cureton, E. E. (1965). Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement*, *25*(2), 327-346.

Folger, T. D., Bostic, J., & Krupa, E. E. (2023). Defining test‐score interpretation, use, and claims: Delphi study for the validity argument. *Educational Measurement: Issues and Practice, 42*(3), 22-38.

Hill, H. C., & Shih, J. C. (2009). Research commentary: Examining the quality of statistical mathematics education research. *Journal for Research in Mathematics Education*, *40*(3), 241-250.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198-211.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.2307/23353796

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

Markus, K. A. (2016). Alternative vocabularies in the test validity literature. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 252-267.

Melhuish, K., & Hicks, M. D. (2020). A validity argument for an undergraduate mathematics concept inventory. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge* (pp. 121-151). Routledge.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 178-197.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.

Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 268-280.

Sireci, S. G. (2016). Comments on valid (and invalid?) commentaries. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 319-321.

Sireci, S., & Padilla, J. L. (2014). Validating assessments: introduction to the special section. *Psicothema*, 97-99.

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

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice, 17*(2), 117-132.