Accepted: 6 October 2023

DOI: 10.1111/ssm.12621

WILEY | SSMA

EDITORIAL

Engaging hearts and minds in assessment and validation research

Assessment continues to be an important conversation point within Science, Technology, Engineering, and Mathematics (STEM) education scholarship and practice (Krupa et al., 2019; National Research Council, 2001). There are guidelines for developing and evaluating assessments (e.g., AERA et al., 2014; Carney et al., 2022; Lavery et al., 2019; Wilson & Wilmot, 2019). There are also *Standards for Educational & Psychological Testing* (*Standards*; AERA et al., 2014) that discuss important relevant frameworks and information about using assessment results and interpretations. Quantitative assessments are used as part of daily STEM instruction, STEM research, and STEM evaluation; therefore, having robust assessments is necessary (National Research Council, 2001). An aim of this editorial is to give readers a few relevant ideas about modern assessment research, some guidance for the use of quantitative assessments, and framing validation and assessment research as equity-forward work.

1 | MODERN ASSESSMENT BACKGROUND

A *test* is a tool that systematically samples a respondent's behavior or knowledge, much like a test of content or pedagogical knowledge for example (AERA et al., 2014). An *assessment* also evaluates a respondent's qualities but does so in a way that might not include correct responses. Surveys, scales, observation protocols, and instruction logs fall into this group of tools. While tests and assessments are different data collection tools, this editorial uses the terms test and assessment interchangeably to promote readability because those terms are frequently used interchangeably across wide bodies of scholarly literature.

Assessments should be grounded in claims and evidence (Kane, 2013; Shepard, 2016), much like a mathematics property is grounded in evidence and reasonable justification (Folger et al., 2023). Claims and evidence form an argument for a test's uses and interpretations (Kane, 2013). Some claims describe the results from an assessment, other claims reflect the development of an assessment for a particular purpose. Claims are supported by validity evidence that has been gathered, as well as scholarly and community wisdom (Folger et al., 2023; Kane, 2013). A test is not valid. Validity is an attribute of how test results are interpreted and used. Validity is a unitary concept that describes the degree to which evidence and theory support a test's results and interpretations (AERA et al., 2014). Claims are supported with evidence related to a validity source: (a) test content, (b) response process, (c) relations to other variables, (d) internal structure, and (e) consequences of testing and bias (AERA et al., 2014; Melhuish & Hicks, 2019). It is not necessary to have evidence from all five validity sources; yet greater and more diverse evidence can support more robust claims (AERA et al., 2014). As the "the stakes for the test taker [increase] and the need to protect the public increase" (AERA et al., 2014, p. 3), then so does the quality and quantity of validity evidence and claims. The outcomes from a study that uses a quantitative instrument are substantially influenced by the instrument's qualities, which reflects the ways an instrument is grounded in robust claims and sound validity evidence (Bostic, 2017).

Historically, test developers have focused on test content and internal structure evidence within mathematics education contexts (Bostic et al., 2021). Such practices omit important evidence that can implicate test results (Bostic, 2019). For example, response process evidence gathered from think alouds can support claims about the degree to which hypothesized respondent's responses align with a test developer's intentions (Bostic, 2021; Folger et al., 2023; Melhuish & Hicks, 2019). Having information about how potential respondents engage with a test confirms that items are created appropriately and elicit a desired response (Bostic, 2021; Padilla & Benitez, 2014). Taken collectively, it is important to maintain a broad perspective during validation research and consider claims and multiple sources of validity evidence.

2 | GUIDANCE FOR THE USE OF QUANTITATIVE ASSESSMENTS

Robust scholarly quantitative results come from high quality tests and appropriate uses of those tests

(Bostic, 2017). Sometimes it can be difficult to discern whether an assessment will fit an intended need. For example, does a test of middle school mathematics achievement serve as a sufficient proxy for measuring middle school students' mathematics problem solving? An achievement test might purport to include problem solving; however, that may be loosely defined or difficult to discern from the item qualities. Thus, a researcher seeking to measure mathematical problem solving might be assessing an entirely different construct using a mathematics achievement measure and potentially generate misleading results. Claims drawn from those results can have longstanding impacts, and potentially misinform future scholarship (Cronbach, 1988). Thus, it is central to associate the measure as closely as possible with the desired construct.

The *Standards* (AERA et al., 2014) recommend that users evaluate tests based upon the following: (a) professional judgment, (b) quality of validity evidence and claims, (c) available alternatives, and (d) applicable guidelines/laws. Conducting literature searches for viable instruments can be time consuming, problematic, and unhelpful. Furthermore, different institutions have differing access to scholarly manuscripts because institutions purchase different journal packages (Bostic et al., 2021). Hence, access to scholarship while searching for assessments can implicate research quality.

A recent National Science Foundation-funded project called Validity and Measurement in Mathematics Education (V-M^{2}Ed) includes numerous scholars seeking to synthesize mathematics education tests used between 2000 and 2020 (NSF#1920619; 1920621). I serve as a Primary Investigator alongside Dr. Erin Krupa, who is also a Primary Investigator. One V-M^{2}Ed goal is to create a repository that is free to users and can be accessed by anyone to search for quantitative mathematics tests and assessments. It will also allow others to add additional tests and assessments to the repository. The repository release date is 2024. A second goal from V-M^{2}Ed is to strengthen a growing community of scholars across many disciplines including but not limited to: mathematics and statistics education, psychometrics, assessment, special education, and others. Nearly 100 scholars across the United States, Canada, and Caribbean have engaged in this project in some form since the 2020 conference or an earlier NSF-funded conference in 2017 (NSF #1644314). Readers interested in this work might explore this website and engage with V-M^{2}Ed presenters at conferences or over email. Exploring the assessments and their associated claims and validity evidence may kickstart research ideas and/or fill a need for those seeking a quantitative test for their scholarship.

3 | VALIDATION RESEARCH AS EQUITY-FORWARD SCHOLARSHIP

Validation research is an important part of assessment scholarship. This work includes developing, critically examining, and revising tests. Cronbach (1988) and Shepard (2016) have reminded assessment scholars of their power within validation scholarship. The interpretation and uses of test results can implicate the rights, privileges, and opportunities for others. Poorly constructed tests with ill-defined or absent test-score interpretation and uses can lead to harmful consequences. This is discussed at length in the *Standards* (AERA et al., 2014), Jonson and Geisinger's (2022) text on fairness and bias, and other peer-reviewed articles (e.g., Shepard, 2016). To that end, everyone engaged in validation research must view their work as equity forward. Assessment scholars have, and should continue to leverage, ways to promote equity within the everchanging educational landscape using their knowledge, wisdom, experiences, and relationships.

4 | MOVING FORWARD

A goal with this editorial was to give readers some brief insight into modern assessment theory, guidance and recommendations for conducting validation research, and begin to frame validation scholarship as equity-forward work. While each of these topics can garner its own editorial or research manuscript, the aim was to give readers a cognitive amuse bouche for future learning. Together, the research and practice-based communities of assessment scholars can work together to promote positive outcomes for others. As an assessment scholar myself, I heed Cronbach's advice and encourage others to do so as well: "...we [assessment experts] do our damnedest—no holds barred—with our minds and our hearts [everyday]" (1988, p. 14).

ACKNOWLEDGMENTS

I would like to gratefully acknowledge feedback and support from Tim Folger, Erin Krupa, and Christie Martin on drafts of this editorial.

FUNDING INFORMATION

National Science Foundation, Grant/Award Numbers: 2201165, 2100988, 1920621, 1720646

Jonathan D. Bostic

Bowling Green State University, Bowling Green, Ohio, USA

Correspondence
Jonathan D. Bostic, Bowling Green State University, Bowling Green, OH, USA.
Email: bosticj@bgsu.edu

Ideas in this work stem from multiple grant-funded research studies supported by the National Science Foundation (NSF# 1720646; 1920621; 2100988; 2201165). Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation.

REFERENCES

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

Bostic, J. (2017). Moving forward: Instruments and opportunities for aligning current practices with testing standards. *Investigations in Mathematics Learning*, *9*(3), 109–110.

Bostic, J. (2019). We can do better! *Intersection Points*, *44*(6), 3–4.

Bostic, J. (2021). Think alouds: Informing scholarship and broadening partnerships through assessment. *Applied Measurement in Education*, *34*(1), 1–9.

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education*, *24*, 5–31. https://doi.org/10.1007/s10857-019-09445-0

Carney, M., 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.

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

Folger, T., Bostic, J., & Krupa, E. (2023). Defining test-score interpretation, use, and claims: Delphi study for the validity argument. *Educational Measurement: Issues and Practice*, *42*(3), 22–38. https://doi.org/10.1111/emip.12569

Jonson, J. L., & Geisinger, K. F. (2022). *Fairness in educational and psychological testing: Examining theoretical, research, practice, and policy implications of the 2014 standards*. AERA.

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

Krupa, E., Carney, M., & Bostic, J. (2019). Approaches to instrument validation. *Applied Measurement in Education*, *32*(1), 1–9.

Lavery, M., Jong, C., Krupa, E., & Bostic, J. (2019). Developing an assessment with validity in mind. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 12–39). Routledge.

Melhuish, K., & Hicks, M. (2019). A validity argument for an undergraduate mathematics concept inventory. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Research instruments and perspectives* (pp. 121–151). Routledge.

National Research Council. (2001). *Knowing what students know. Committee on the foundations of assessment*. National Academies Press.

Padilla, J.-L., & Benitez, I. (2014). Validity evidence based on response process. *Psichotherma*, *26*(1), 136–144.

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

Wilson, M., & Wilmot, D. (2019). Gathering validity evidence using the BEAR assessment system (BAS): A mathematics assessment perspective. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 63–89). Routledge.