GATHERING VALIDITY EVIDENCE TO SUPPORT MATHEMATICS EDUCATION SCHOLARSHIP

Jonathan D. Bostic
Bowling Green State Univ.
bosticj@bgsu.edu

Erin Krupa
North Carolina State Univ.
eekrupa@ncsu.edu

Timothy Folger
Bowling Green State Univ.
tdfolge@bgsu.edu

Brianna Bentley
North Carolina State Univ.
blbentle@ncsu.edu

David Stokes
North Carolina State Univ.
djstokes@ncsu.edu

Validity and validation is central to conducting high quality quantitative mathematics education scholarship. This presentation aims to support scholars engaged in quantitative research by providing information about the degrees to which validity evidence related to their instrument use or interpretation, were found in mathematics education scholarship. Findings have potential to steer future quantitatively focused scholarship and support equity aims.

Keywords: assessment, research methods

The inferences and interpretations drawn from quantitative assessments are largely grounded in the validity evidence and their associated claims (AERA et al., 2014; Carney et al., in press; Kane, 2016). Mathematics education scholarship using quantitative assessments has not consistently adhered to strong validity and validation practices as seen by the limited presence of validity evidence related to many instruments used with K-12 students and teachers (Bostic et al., 2019, Bostic et al., 2021; Krupa et al., 2019). A purpose for the present study is to serve as an educative piece for mathematics educators who intend to develop or use quantitative assessments in their research. We offer results about the degree to which validity has been taken up in mathematics education scholarship that uses quantitative instruments.

Related Literature

The *Standards for Educational and Psychological Testing* ([*Standards*] AERA et al., 2014) describe five sources for validity: test content, response process, relations to other variables, internal structure, and consequences from testing. Reliability is a related but not sufficient condition of validity (AERA et al., 2014; Kane, 2012). While the *Standards* describe some approaches for each source, those descriptions are not intended to be exhaustive. A special issue of *Psicotherma* (2014, 26(1)) includes articles related to each of those sources. Here again, the authors of those articles indicate that their description of each validity source is to introduce readers to that source and are not intended to be comprehensive. Thus, there is a need for a more comprehensive list of data collection approaches related to each validity source.

Mathematics education scholarship has started to address validity within the context of quantitative research across K-20 students, as well as preservice and inservice teacher settings (e.g., Bostic & Sondergeld, 2015; Carney et al., 2017; Gleason et al., 2019; Hill & Shih, 2009; Melhuish & Hicks, 2019; Walkowiak et al., 2014; Wilhelm & Berebitsky, 2019). These authors provide discussions about how they explored validity evidence and serve as potential roadmaps for others doing validation work. There are some common approaches to gathering validity evidence. The present study intends to summarize literature about approaches to gathering validity evidence so that quantitative assessment developers, users, and reviewers have a more comprehensive understanding of what has been done previously, and what approaches are viable.

Our research question was: In what ways is validity evidence and arguments present in mathematics education scholarship that uses quantitative instruments between 2000-2020?

Methods

Context for study

In 2019, 39 mathematics education scholars convened at a 2-day conference led by the authors. Scholars included mathematics educators, psychometricians, special educators, and policy experts who had previously conducted quantitative mathematics or statistics education assessment work. Conference leaders asked attendees to form small groups and brainstorm viable data collection approaches that might generate validity evidence for that source. After sharing ideas, groups rotated to each approach and left feedback, followed by whole-group discussions. The product was an extensive list with at least five unique approaches to gathering validity evidence for each validity source. There were still questions about whether there were other approaches and how to define some of these approaches for a broad audience (e.g., factor analysis). To that end, the authors of this submission reviewed literature and sought definitions to create an Evidence Types Guidebook. The definitions were sent to conference attendees for feedback and revised as needed. Independent of that work, conference attendees have been conducting syntheses of literature across a variety of contexts including teacher education, elementary and secondary students, and statistics education.

An Evidence Types Guidebook served to facilitate the identification of approaches typically utilized within a validity evidence type. For each of the evidence types (i.e., test content, response processes, internal structure, relation to other variables, and consequences of testing) within the Guidebook, a general definition of the type was given, followed by a list of methods commonly used to support a validity claim within the respective category. For example, the internal structure section included approaches such as factor analysis, item response theory, latent class analysis, and other approaches commonly used to assess and support claims of validity related to the internal structure of a quantitative instrument. Each of these methods also contained definitions and citations for further exploration and information. The Guidebook content was aligned with the validity evidence repository framework and served to support participants, in general, throughout the framework application process.

Data Sources and Analysis

Our data collection and analysis process is summarized here; more details are provided in Bostic et al. (2022). The PRISMA statement guided the literature search (Rethlefsen et al., 2021). The top 24 mathematics education journals (Williams & Latham, 2017) were searched for studies using quantitative instruments. Articles that included quantitative instruments were culled to create a list of instruments. Next, validity evidence was sought for each instrument through a literature search using google scholar. Instrument names and keywords were used to generate an appropriate sample space. As an example from teacher education, over 3,000 articles were examined, which led to over 300 instruments Evidence was coded as being connected to a validity source or reliability. Validity claims in support of arguments were also coded. We share results from those syntheses as a means to illuminate the frequencies of various approaches as well as opportunities for use of new approaches.

Findings

Overall, synthesis groups searched for validity evidence of 190 instruments and found 278 articles with descriptions of validity evidence (see Table 1). The majority of articles (83%) did not contain an interpretation statement or a use statement. In addition, 73% of the articles that

contained validity evidence did not specify any claims. An example of a use statement comes from the Statistics Education synthesis group: "The availability of an instrument such as the Attitudes Toward Research (ATR) scale which has been designed for students, may provide information concerning motivational aspects associated with learning research, and might also have potential for identifying distinctive attitude profiles of students who find research problematic. Overall however, this study's results validate the utility of the ATR scale in measuring student attitudes toward research" (Papanastasiou, 2005, p. 23).

Table 1. Instrument and Article Overview

	Elementary (K - 6) Tests & Instruments	Secondary (7 - 12) Tests & Instruments	Statistics (K - 20) Tests & Instruments	Teacher Education Instruments	Combined
Number of instruments	59	27	16	88	190
Number of articles	92	36	52	98	278
Articles with an interpretation statement	17 18.48%	4 11.11%	19 36.54%	7 7.14%	47 16.91%
Articles with a use statement	18 19.57%	2 5.56%	20 38.46%	2 2.04%	42 15.11%
Articles with a claim	37 40.22%	6 16.67%	15 28.85%	15 15.31%	73 26.26%

Considering the distribution of the five types of validity as well as reliability: Internal structure, reliability, and test content were the most frequently located. Table 2 shows the frequency of each evidence type across different areas. The most frequently used method for each evidence type is displayed in Table 3. Some frequencies are quite high (e.g., alignment with frameworks) whereas the mode for other validity sources was quite low (e.g., quantitative DIF analysis).

Table 2. Evidence Type Frequency

	Elementary (K - 6) Tests & Instruments	Secondary (7 - 12) Tests & Instruments	Statistics (K - 20) Tests & Instruments	Teacher Education Instruments	Combined
Consequences of Testing	4 1.73%	0 0.00%	5 2.81%	1 0.69%	10 1.71%
Internal Structure	48 20.78%	7 21.88%	34 19.10%	29 20.14%	118 20.17%
Relations to Other Variables	33 14.29%	1 3.13%	27 15.17%	19 13.19%	80 13.68%
Reliability	68 29.44%	9 28.13%	29 16.29%	53 36.81%	159 27.18%

Response Process	27 11.69%	3 9.38%	13 7.30%	6 4.17%	49 8.38%
Test Content	51 22.08%	12 37.50%	70 39.33%	36 25.00%	169 28.89%
Total Number of Evidence Types	231	32	178	144	585

Table 3. Mode of approach for each validity source and reliability

Validity Source or Reliability	Most common type of approach	(count, %)
Test content	Alignment with frameworks	n = 45, 26%
Response Process	Student written work	n = 19, 38%
Relations to Other Variables	Correlation analysis	n = 41, 51%
Internal Structure	Confirmatory factor analysis	n = 41, 34%
Consequences of Testing	Item functioning such as DIF	n = 3, 30%
Reliability	Internal consistency, alpha	n = 19, 38%

Discussion

It is clear from analyzing the validity evidence from this sample of instruments that modern notions of validity and validation arguments (AERA et al., 2014; Author, in press; Kane, 2016), have not necessarily been taken up by the field. We do not blame authors for this omission. It may be that validity evidence is removed during the editing process. Authors may not be prepared to conduct validation work. It has also been shown that 75% of mathematics education graduates take two or less quantitative research courses, where validity and validation might be discussed (Shih et al., 2019). Few instruments presented in existing scholarship are accompanied with an explicit statement describing the intended interpretation and use of test scores. Further, we found little validity evidence based on consequences of testing and response processes. Given current equity issues in mathematics education, it is a concern that there is not more evidence of consequences from testing and bias, especially to ensure fair use of the score interpretations from the tests.

Validity is naturally an equity issue (AERA et al., 2014; Cronbach, 1988). Otherwise, tests may have bias and test scores may be used unfairly. Cronbach (1988) proclaimed, “Tests that impinge on the rights and life chances of individuals are inherently disputable” (p. 6). Furthermore, the inferences drawn from tests that lack a validity argument may not be accurate (Carney et al., in press). To yield accurate inferences about student learning or teacher practice, it is critical for scholars to have tests and instruments with strong validity evidence and robust validity arguments.

Acknowledgments

This research is funded in part by grants from the National Science Foundation (1920619, 1920621). All findings are those of the authors and not necessarily of the funding agency.

References

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.

Bostic, J.D., Carney, M., Casey, S, Engledowl, C., Folger, T., Gallagher, M., Howell, H., Smith, W., Tjoe, H., & Wilhelm, A. (2022, February). *Choose your instruments wisely: Supporting mathematics teacher educators' research and practice.* Symposium presented at annual Association of Mathematics Teacher Educators Conference. Henderson, NV.

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.

Bostic, J., Krupa, E., Carney, M., & Shih, J. (2019). Reflecting on the past and thinking ahead in the measurement of students' outcomes. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 205-229). New York, NY: Routledge.

Bostic, J., & Sondergeld, T. (2015). Measuring sixth-grade students' problem solving: Validating an instrument addressing the mathematics Common Core. *School Science and Mathematics Journal, 115, 281-291*.

Carney, M., Bostic, J., Krupa, E., & Shih, J. (in press). Instruments and use statements for instruments in mathematics education. *Journal for Research in Mathematics Education*. Accepted for publication.

Carney, M., Cavey, L., & Hughes, G. (2017). Assessing teacher attentiveness: Validity claims and evidence. *Elementary School Journal, 118*(2), 281-309.

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

Gleason, J., Livers, S., & Zelkowski, J. (2017) Mathematics Classroom Observation Protocol for Practices (MCOP2): A validation study. *Investigations in Mathematics Learning, 9*(3), 111-129.

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

Kane, M.T. (2012). All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 66-70.

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development. (2nd ed., pp. 64-80).* Routledge.

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

Melhuish, K., & Hicks, M. (2019). Title redacted for blinding. In In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 121-151). New York, NY: Routledge.

Papanastasiou, E. C. (2005). Factor Structure of the "Attitudes toward Research" Scale. *Statistics Education Research Journal, 4*(1), 16–26.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic reviews*, *10*(1), 1-19.

Shih, J., Reys, R., Reys, B., & Engledowl, C. (2019). A profile of mathematics education doctoral graduates' background and preparation in the United States. *Investigations in Mathematics Learning, 11(*1), 16-28.

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics, 85*(1), 109-128.

Wilhelm, A. G., & Berebitsky, D. (2019) Validation of the mathematics teachers' sense of efficacy scale. *Investigations in Mathematics Learning, 11*(1), 29-43.

Williams, S. R., & Leatham, K. R. (2017). Journal quality in mathematics education. *Journal for Research in Mathematics Education, 48*(4), 369-396.