DOI: 10.1111/ssm.12410

WILEY | SSMA

EDITORIAL

Shifting sights on STEM education quantitative instrumentation development: The importance of moving validity evidence to the forefront rather than a footnote

One would be hard pressed to complete a graduate program in education and not, at a minimum, hear about the concepts of validity (measuring what we intend to measure) and reliability (measuring consistently). As educational researchers, we are well aware that strong validity and reliability evidence for our instruments are fundamental for the advancement of research. Even so, hard scientists often think less of educational research because "the social sciences are, well, 'soft', and lacking methodological rigour" ("In praise of soft science," 2005, p. 2). This criticism is in part due to the nature of the constructs under study. While our hard science colleagues are investigating scientific phenomena with well-established tools to measure outcomes such as growth in height or differences in speed; social science researchers are attempting to quantify or explain constructs that are far more challenging to measure such as human attitudes and beliefs, or cognitive abilities. The challenge may also in part be due to a lack of measurement training (Liu, 2010; Shih, Reys, Reys, & Engledowl, 2019; Smith, Conrad, Chang, & Piazza, 2002), which is essential for social scientists to develop and validate sound instruments.

To address this inherent concern about instrumentation rigor in the social sciences, *The Standards for Educational and Psychological Testing* were first released in 1966 by a collaboration comprised of American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). In 2014, the most recent version of *The Standards* (AERA, APA, & NCME, 2014) discussed the need to collect, evaluate, and document multiple forms of validity evidence for the results and interpretations of a quantitative instrument to be judged suitable for a specified intent. Further, with greater validity evidence to support the validity argument of an instrument, stronger inferences may be drawn regarding instrumentation soundness (AERA et al., 2014; Kane, 2016).

While there are numerous forms of validity evidence discussed within volumes of literature, there are five specific types *The Standards* urge developers of educational and psychological assessments to evaluate: test content, response processes, internal structure, relationship to other variables, and consequential (AERA et al., 2014). Test content validity evidence investigates instrument item alignment (test content) with the construct to be measured (theoretical trait). Supporting evidence often comes from subject matter experts (SMEs) evaluating item-to-construct alignment and can be logical or empirical (qualitative) (Sireci & Faulkner-Bond, 2014). Response process validity evidence appraises participant responses or performance alignment with the test construct (Leighton, 2017). Generally, data to support response process validity is qualitative and collected through cognitive interviews, think alouds, or focus group interviews with a sample of typical respondents to check that they understand items and respond in ways developers envisioned (Padilla & Benitez, 2014). Internal structure validity evidence is assessed through psychometric methods to explore: (a) instrument dimensionality, (b) measurement invariance, and (c) instrument reliability (Rios & Wells, 2014). Relationship to other variables validity evidence often uses statistical testing to investigate instrument outcome associations with other variables hypothesized to be related (either positively or negatively) (Beckman, Cook, & Mandrekar, 2005). Consequential validity evidence and bias are often collected through qualitative data to examine how participants perceive the assessment to have impacted them or how the results could have impacted participants (Bostic & Sondergeld, 2015), but such data can be examined quantitatively too. Even though each of the five sources of validity evidence are considered important, it is significant to note that research examining beyond content and internal structure of instruments is rare (Beckman et al., 2005) (see as exceptions Bostic, Matney, & Sondergeld, 2019; Bostic & Sondergeld, 2015).

Supporting the value of quantitative instrument validity and reliability evidence in the social sciences, prominent federal agencies such as the National Science Foundation (NSF) and Institute of Educational Sciences (IES) regularly request grant proposals centered on the rigorous development of new tools and assessments. One such collaborative mathematics project currently funded through NSF's DRK-12 program is entitled *Developing and Evaluating Assessments of Problem Solving* (*DEAP*).[1] This four-year project started in 2017 and focuses on rigorously developing and collecting extensive

[1]NSF Award Number: 1,720,646 – Bostic (PI) & Matney (Co-PI) from BGSU; Sondergeld (PI) from Drexel.

validity evidence for three Problem Solving Measures (*PSM3*, *PSM4*, *PSM5*) in grades 3 to 5. A team of mathematics educators and psychometricians work on this initiative to ensure assessments are aligned with Common Core Standards, vertically equated to each other, and linked to prior developed middle grades *PSM*s in grades 6 to 8. According to the NSF abstract, "this project fills a need in the field as no set of measures uses vertical equating to assess elementary students' problem-solving performance in a rigorous fashion within the contest of state testing" (NSF, n.d., p. 2). In a somewhat similar vein, NSF also funded two rounds of a collaborative research project entitled *Validity Evidence for Measurement in Mathematics Education* (*V-M²ED*).[2] The purpose of the current initiative is not to develop new instruments, but to create criteria for evaluating validity evidence of currently used quantitative mathematics instruments and synthesize validity evidence from published literature. Teams of mathematics educators, psychometricians, statisticians, policy experts, and graduate students are collaborating to complete *V-M²ED* efforts. These are just two samples that underscore the importance of extensive STEM educational instrumentation development, validation, and publishing for external review and examination.

With support for research and reporting on instrumentation soundness by many of education's leading organizations and funding agencies, one might expect validity studies to be common. For this piece, a review of the top 20 mathematics, science, and/or STEM education journals (according to Scimago Journal Rankings https://www.scimagojr.com/) was conducted to investigate this hypothesis. Journal editorial staff were emailed and asked if they would consider publishing validation studies. While 40% of journal editors did not reply, 20% ($n = 4$) indicated they would not accept a validation study and 40% ($n = 8$) said they would if it aligned with their journal's aims and goals. From the 16 journals that either did not respond or reported they would consider validation studies, a deeper dive into the number of research reports and number of instrumentation validity manuscripts published over the last 5 years was completed. Less than 2% of empirically based publications in these journals were mathematics, science, or STEM instrument validity studies.

If top science, technology, engineering, and mathematics (STEM) education journals are willing to accept instrument validity studies, and the field deems this as an important venture, then why are there so few disseminated through peer-reviewed publications? Perhaps the answer is partially due to time or expertise. As established through both the *DEAP* and *V-M²ED* projects, validation work takes significant time (multiple years) and is conducted through an iterative process as collaborative efforts between experts of various sorts (SMEs and psychometricians). While it is important to note that many quantitative mathematics, science, and/or STEM education outcome studies do provide reliability statistics (e.g., Cronbach alpha) and/or mention the instrument's outcomes have been demonstrated to be a valid indicator of the content being assessed, this is no substitute for a rigorously developed and published validity study. If it is a goal of researchers in the fields of mathematics, science, and STEM education to have their quantitative study findings viewed as scientific among broader audiences, the manner in which instruments are designed, tested, and disseminated must be moved to the forefront rather than placed in a footnote. This shift is necessary to advance the field and allow for valid and reliable outcomes research to be produced from meaningful measures.

[2]NSF Award Number: 1,644,314 – Bostic (PI) from BGSU & Krupa (PI) from NC State.

Toni A. Sondergeld

Drexel University, School of Education, Philadelphia, PA, USA

Drexel University, School of Education, 3401 Market St, 3rd Floor, Philadelphia, PA 19104, USA.
Email: tas365@drexel.edu

ORCID

Toni A. Sondergeld https://orcid.org/0000-0001-7264-5607

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of General Internal Medicine*, *20*(12), 1159–1164. https://doi.org/10.1111/j.1525-1497.2005.0258.x

Bostic, J. D., Matney, G., & Sondergeld, T. A. (2019). A validation process for observation protocols: Using the revised SMP look-for protocol as a lens on teachers' promotion of the standards. *Investigations in Mathematics Learning*, *11*(1), 69–82.

Bostic, J. D., & Sondergeld, T. A. (2015). Measuring sixth-grade students' problem-solving: Validating an instrument addressing the mathematics common core. *School Science and Mathematics Journal*, *115*(6), 281–291. https://doi.org/10.1111/ssm.12130

In praise of soft science. (2005). *Nature*, *435*(1003). https://doi.org/10.1038/4351003a

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(3), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Leighton, J. P. (2017). *Using think aloud interviews and cognitive labs in educational research*. Oxford, UK: Oxford University Press.

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach*. Charlotte, NC: Information Age.

National Science Foundation (NSF) (n.d.). *Award abstract #1720646: Collaborative research: Developing & evaluating assessments of problem solving*. Retrieved from https://www.nsf.gov/awardsearch/showAward?AWD_ID=1720646

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

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108–116.

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. https://doi.org/10.1080/19477503.2017.1375357

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107.

Smith, E., Conrad, K., Chang, K., & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. *Journal of Nursing Measurement*, *10*, 189–206. https://doi.org/10.1891/jnum.10.3.189.52562