Journal for Research in Mathematics Education

The Journal for Research in Mathematics Education is an official journal of the National Council of Teachers of Mathematics (NCTM). *JRME* is the premier research journal in mathematics education and is devoted to the interests of teachers and researchers at all levels—preschool through college.

ARTICLE TITLE:

Interpretation and Use Statements for Instruments in Mathematics

AUTHOR NAMES:

Carney, Michele B.; Bostic, Jonathan; Krupa, Erin; Shih, Jeff

DIGITAL OBJECT IDENTIFIER: 10.5951/jresematheduc-2020-0087

VOLUME: 53

ISSUE NUMBER: 4

Mission Statement

The National Council of Teachers of Mathematics advocates for high-quality mathematics teaching and learning for each and every student.

CONTACT: jrme@nctm.org

Journal for Research in Mathematics Education
2022, Vol. 53, No. 4, 334–340

Research Commentary

Interpretation and Use Statements for Instruments in Mathematics Education

Michele B. Carney
Boise State University

Jonathan Bostic
Bowling Green State University

Erin Krupa
North Carolina State University

Jeff Shih
University of Nevada–Las Vegas

This Research Commentary addresses the need for an instrument abstract—termed an Interpretation and Use Statement (IUS)—to be included when mathematics educators present instruments for use by others in journal articles and other communication venues (e.g., websites and administration manuals). We begin with presenting the need for IUSs, including the importance of a focus on interpretation and use. We then propose a set of elements—identified by a group of mathematics education researchers, instrument developers, and psychometricians—to be included in the IUS. We describe the development process, the recommended elements for inclusion, and two example IUSs. Last, we present why IUSs have the potential to benefit end users and the field of mathematics education.

Keywords: Tests; Measure development and validation; Instrumentation theory

The purpose of this Commentary is to increase discussion in the field of mathematics education around conceptualizations of validity and instrument validation and to describe the need for instrument abstracts—termed *Interpretation and Use Statements* (IUSs)—and the elements such abstracts could incorporate. To begin, we briefly describe the issues the IUS is designed to address through the lenses of (a) an individual considering use of an instrument in a mathematics education context and (b) the broader field of research on mathematics education.

Identifying instruments for use in an individual's particular context can be difficult. First, instruments typically have multiple users (e.g., mathematics education researchers, educators, graduate students, district personnel), and these users vary greatly with respect to their expertise in finding and evaluating the evidence associated with validation. Second, whereas excellent resources, such as the *Mental Measurements Yearbook* (Carlson et al., 2017) are available for commercial instruments, no comprehensive repository of instruments focused on important constructs in mathematics education currently exists, which often makes it difficult to find instruments. Third, when researchers rely on journal searches, parsing through the details of an academic article to determine whether an instrument is appropriate for your use can be difficult. Having an abstract of the sort we describe here could assist individuals to make a quick initial judgment about the potential usefulness of an instrument and could help ensure informed decisions around score interpretation and use. Regarding the need for IUSs in the context of the broader field of mathematics education research, as we will review in the next section, the field of measure validation has shifted its conceptualization of validity and validation with ramifications for research on mathematics education (Bostic, Krupa, et al., 2019).

In this Commentary, our intent is to foster discussion within the mathematics education community about communicating important information that centers around the stated interpretation(s) and use(s) for the instrument. In advocating for IUSs, we will articulate the elements we are proposing for an IUS along with their descriptions, and provide example IUSs.

This work was supported by Grants No. 1644314, 1644321, 1920619, and 1920621 from the National Science Foundation. Any opinions expressed in this manuscript are those of the authors and do not reflect the views of the National Science Foundation. The second through fourth authors contributed to this work equally and are listed in alphabetical order.

Shifts in Conceptualizations of Instrument[1] Validity and Validation

Validity and validation literature has shifted to position validity as a property of score interpretations for proposed uses, and not a property of the instrument itself—that is, tests are not valid (American Educational Research Association [AERA] et al., 2014;[2] Kane, 2013). This focus is intended to ensure that assessment results are not interpreted or used in inappropriate ways simply because an instrument is deemed valid and reliable. The *Standards for Educational and Psychological Testing* (AERA et al., 2014; henceforth the *Standards*), which represent more than 60 years of consensus among measurement professionals (Plake & Wise, 2014), position validation "as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (AERA et al., 2014, p. 11). Therefore, validation is focused on the specified interpretation and use. When an instrument user wants to interpret or use scores in a manner different from what is specified by the developer and the associated validation argument, additional validation must be conducted to ensure that is appropriate.

The conceptualization of argument-based validation as providing evidence in relation to the interpretation and use of test scores is clearly articulated in the *Standards*, but it marks a substantial shift in understanding and practice. Examinations and reviews of the education literature have indicated that argument-based validation is not common in practice (Cizek et al., 2008, 2010; Wolming & Wikström, 2010). Similarly, reviews have indicated a general lack of focus on validity in mathematics education (Bostic et al., 2021; Hill & Shih, 2009; Minner et al., 2013). When validity is mentioned, the focus tends to be on presenting statistics (e.g., Cronbach's alpha or factor analysis) exclusive of evidence that the "instruments tap what they claim to" (Hill & Shih, 2009, p. 248). A consistent lack of argument-based approaches to validity continues to exist within mathematics education scholarship (Bostic, Krupa, et al., 2019). A brief review of the literature associated with instruments in the *Compendium of Research Instruments for STEM Education* (Minner et al., 2013) indicated that many instrument developers tend to still focus on validity as a property of an instrument as opposed to a property of score interpretations for proposed uses (e.g., Barker & Ansorge, 2007). Taken collectively, these findings indicate that the field of education research, including mathematics education, may find it useful to have resources highlighting conceptualizations of validity and validation that focus on the interpretation and use of instrument scores.

A New Kind of Abstract: The Interpretation and Use Statement

We argue that our field would benefit from the development of a new kind of abstract to address the need to provide explicit information to end users on the intended score interpretation for a proposed use of an instrument. Additional elements (e.g., target population and context for administration) are also included to help end users make an initial determination of whether an instrument is likely to be appropriate for their use situation. The abstract we are proposing, the IUS, is not intended to take the place of an extensive description of the development process and its associated validity argument and evidence. Instead, it provides an initial evaluation point for the end user, as well as an explicit statement of the intended interpretations for proposed uses to support development or examination of a validity argument. The IUS should be presented when a significant focus of a journal article or other communication venue is presenting an instrument or its associated validity evidence. In other words, it should be presented when and where end users are likely to interact with the instrument (e.g., websites, instrument manuals, or journal articles).

Developing Recommendations for IUSs

The National Science Foundation funded the Validity Evidence for Measurement in Mathematics Education (V-M^{2}Ed) conference to focus on bringing together "researchers . . . from different theoretical and methodological perspectives to contextualize current conceptions of validity within the field of mathematics education" with the purpose of "strengthening quantitative measure quality in mathematics education, with a specific focus on validity" (V-M^{2}Ed, n.d., para. 2). One goal of the conference was to develop a set of recommended elements that should be clearly and succinctly stated by instrument developers in the form of a statement to end users about the intended score interpretations for proposed uses[3] (i.e., the elements that make up the IUS).

---

[1] The *Standards* describe a *test* as "a device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process," whether responses are "evaluated for their correctness or quality" or "used for measures of attitudes, interest, and dispositions" (AERA et al., 2014, p. 2). Within mathematics education scholarship, the language convention has been to use "test" more narrowly to characterize instruments for which responses are evaluated for correctness or quality. To that end, we intentionally use the term "instrument" throughout the manuscript as a broad term that includes tests, surveys, questionnaires, observation protocols, and others.

[2] Although we refer to the most recent edition of the *Standards* (AERA et al., 2014), the previous edition of the *Standards* is similar and supports this Commentary equally well.

[3] We wish to explicitly thank the conference participants for their thoughtful contributions to this work.

The 39 participants at the April 2017 meeting were individuals from both higher education (graduate students and faculty) and industry who had an interest in validity. Participants had diverse backgrounds, with expertise in mathematics education, mathematics, psychometrics, or applied measurement and who, aside from the six graduate students, all had terminal degrees in their field. (See https://bit.ly/vmed2017 for information on attendees' roles, expertise, and affiliations.)

The V-M^{2}Ed conference was held over 2 days. On Day 1, one of three primary conference goals was the identification of a set of recommendations for the elements to include in an IUS. A set of elements generated by the conference leaders (the authors of the present Commentary), drawing on aspects highlighted in the *Standards* and in the research literature and written as questions, was given to conference participants. Using these elements as a starting point, attendees were asked to draft an example IUS for a measure around a construct of their choice. From this work, they were asked to note elements to include and eliminate from the provided list, and to add other elements to discuss for inclusion. This small-group work time was followed by a whole-group discussion on the common elements to include in IUSs. These small- and whole-group recommendations were incorporated in the elements/questions list, and the document was further expanded to provide a draft description of each element/question.

On Day 2, the revised document was used by groups of participants to draft a new example IUS and to give feedback in the form of edits, suggestions, or comments on the elements in the revised document. A whole-group discussion was held and IUS element suggestions were solicited. The small-group notes, example IUSs, and field notes from the whole-group discussion were analyzed following the conference and used to craft the following set of reporting recommendations for IUS elements.

Purpose of the IUS

The IUS, as designed by participants in this conference, is written for the end user to assist in making an initial determination of whether the score interpretation resulting from an instrument aligns with their intended use. The purpose is not to provide detailed technical evidence with which to evaluate the validity of the intended score interpretation. The statement should be succinct and devoid of technical jargon, yet should give the end user enough information to determine whether they should further investigate use of the instrument for their context. The 10 elements described in the next section are recommended for inclusion in an IUS for mathematics education instruments. These recommendations may not necessarily be inclusive of all the elements that should be provided for a particular instrument, and not all elements may be necessary for every instrument. Figure 1 gives an overview of the 10 elements in relation to three overarching categories of construct articulation, operationalization and administration, and scores and usage.

Proposed Elements for an IUS

Element 1: What Construct Is Being Measured?

The theoretical construct, that is "the concept or characteristic that a test is designed to measure" (AERA et al., 2014, p. 11), should be explicitly described. This includes reference to the theory or framework(s) instrument developers used, any

Figure 1

The 10 IUS Elements Organized Into Three Overarching Categories

Construct Articulation	Element 1: What construct is being measured? Element 2: Why is the construct important to measure?
Operationalization & Administration	Element 3: How is the construct measured? Element 4: Who is the target population? Element 5: What is the context for administration? Element 6: What costs are associated with the instrument?
Scores & Usage	Element 7: How are scores calculated? Element 8: How should scores be interpreted? Element 9: How should scores be used? Element 10: What cautions or warnings should be considered?

subconstructs, and expected relations among the subconstructs. Explicitly stating this information for end users assists them in determining whether the conceptualization of the construct aligns with their conceptualization and use scenario.

Element 2: Why Is the Construct Important to Measure?

Is the construct an important educational outcome itself or related to other important educational outcomes? For example, interventions (e.g., professional development) often seek to have an impact on education outcomes (e.g., student mathematics achievement) that are relatively distal to the intervention itself. Explaining why a construct is important to measure often involves connecting a construct that is relatively proximal to the intervention (e.g., teacher self-efficacy) to these more distal outcomes. Describing relationship(s) among constructs or variables helps end users understand how the instrument results connect to broader educational outcomes.

Element 3: How Is the Construct Measured?

The type of instrument should be described (e.g., observation protocol, survey, exam, etc.). This includes the number of items, number of instrument forms available, and item format(s). Constructs are necessarily narrowed through measurement (i.e., operationalization in an instrument). The description of item formats, such as multiple choice and constructed response, provides insight to end users into how a construct was operationalized.

Element 4: Who Is the Target Population?

The group(s) from which data can appropriately be collected should be clearly identified, as well as groups for which use is not appropriate, particularly in situations in which the developer thinks an end user may want to use the instrument with a population other than the one initially targeted. Specifying the target population for end users will allow them to evaluate whether the instrument is appropriate for their use situation. Delineating between populations in which use is appropriate and not appropriate could be particularly helpful.

Element 5: What Is the Context for Administration?

Descriptions of the context for administration may include elements such as the typical setting for administration or data collection, time limits, format for presentation of the instrument (e.g., online or on paper), and so forth. Specifying commonly allowed accommodations may be appropriate. This will help the end user make an initial determination about whether their context for administration aligns well with the intent of the instrument developer.

Element 6: What Costs Are Associated With the Instrument?

Descriptions of cost should consider both time (e.g., for test-taker or administrator) and monetary costs (e.g., instrument access, training to use the instrument, or scoring). Providing this information can save the end user from pursuing an assessment they cannot afford.

Element 7: How Are Scores Calculated?

Developers should clearly indicate the type of instrument score reported (e.g., total raw score, percentile score, or scaled score). Further, who scores the instrument should be clear, as well as how the final score is reached. For example, instruments may require trained scorers for valid score interpretations, resulting in a cost to end users. Describing how scores are calculated provides end users with some understanding of the technical skills, costs, and effort required to calculate scores.

Element 8: How Should Scores Be Interpreted?

Explaining score interpretation involves describing the scope and extent to which the construct is operationalized by the instrument. This is achieved by "delineating the aspects of the construct that are to be represented" (AERA et al., 2014, p. 11). In addition, ideally this element involves substantive, qualitative descriptions of the operationalized construct along a quantitative (i.e., score) continuum (Shepard, 2018). In other words, what words should be used to describe the meaning of a low versus a high numeric score? Last, the score interpretation guidelines should also include (if applicable) whether scores are interpreted in relation to criteria or norms, who interprets the scores, and who makes use of the score interpretations. Providing this information to potential end users is important so they understand what interpretations are possible and appropriate for scores from an instrument.

Element 9: How Should Scores Be Used?

Developers should describe the intended score use(s) in relation to the interpretation. Uses that have clear supporting evidence should be specified. Anticipated uses that are likely to occur but do not have supporting evidence

(i.e., so-called off-label uses) should also be highlighted for end users, including the current lack of evidence to support that use. If appropriate, the intended consequences of score use should be specified. In cases in which the instrument developer anticipates inappropriate uses that may have adverse consequences, the instrument IUS should caution end users against the unsupported uses (see Element 10). Clearly specifying the intended use and nonuse scenarios for end users provides them with an initial judgment of whether the instrument is likely to be useful for their assessment situation. It also has the potential to curb inappropriate uses that may have unintended harmful consequences.

Element 10: What Cautions or Warnings Should Be Considered?

The instrument developer should identify and report on common issues that are likely to result in misinterpretation of scores or unintended consequences, with a strong focus on equity. For example, the end user should be cautioned against using the instrument outside the recommended context or target population in cases in which use outside of that scope is likely to occur. Similarly, the IUS should indicate whether developers recommend or require extensive training to ensure appropriate administration, scoring, or interpretation of an instrument. Explicitly stating likely misinterpretations of scores or unintended consequences can help clarify for the end user that those should be avoided and help solidify their understanding of appropriate interpretation(s) and use(s).

IUS Exemplars

The following example IUSs address each of the elements previously described for a specific instrument relevant to mathematics education. To highlight alignment between the IUS elements and the exemplars, we include parenthetical references to the corresponding element numbers after each relevant portion of the exemplar narratives.

Diagnostic Assessments of Proportional Reasoning

IUS

The Diagnostic Assessments of Proportional Reasoning (DAPR) measure students' composed unit and multiplicative comparison conceptions (Lobato & Ellis, 2010) in proportional reasoning situations (Carney et al., 2019; E1). Composed unit and multiplicative comparison understanding is crucial for conceptual understanding of rate-of-change situations in upper level mathematics and science topics (E2). Understanding a student's location along a hypothetical learning trajectory will assist teachers in targeting classroom instruction to scaffold student learning. The DAPR content is targeted at the standards for middle grades in the Common Core State Standards for Mathematics (National Governors Association Center for Best Practices [NGO Center] & Council of Chief State School Officers [CCSSO], 2010) and has been vetted with students in Grades 6–9 (E4). The DAPR are 20-item fill-in-the-blank assessments available in three equated forms (E3). Hardcopy forms are administered by classroom teachers who impose a 20-min time limit, and students are not allowed access to calculators (E5). The forms are scored as the sum of the number correct, and results are interpreted by the classroom teacher (E7). A student's DAPR score can be interpreted in relation to the hypothetical learning trajectory of composed unit and multiplicative comparison understanding found in Carney and Smith (2017; E8). The scores can be used by classroom teachers to identify instructional scaffolds for students and could be used as one of multiple measures to identify students in need of remediation (E9). Although the items can be aligned to addressing particular standards, the scores should not be used as a comprehensive measure of proportional reasoning or in isolation to identify students in need of remediation (E10). The instruments and user manual are open source (E6).

Commentary

The DAPR were developed out of a need for instruments in which the items were easy to administer and score, and of which teachers could use the results to determine where students were along a trajectory of composed unit and multiplicative comparison conceptions to inform their instruction. However, teachers often mention their plan to use the DAPR as their end-of-unit test for proportional reasoning, which would mean interpreting the results as identifying whether students have mastered grade-level standards. However, the results of the DAPR cannot be used in this way because middle-grades standards for proportional reasoning are typically much broader in focus than the DAPR (in other words, the construct is significantly underrepresented). The DAPR's IUS makes clear the correct interpretation and use and also provides a clear caution that interpreting in relation to grade-level standards is inappropriate. In this way, an end user who may need a comprehensive assessment of proportional reasoning can quickly determine that this instrument is not appropriate for their use.

Revised Standards for Mathematical Practice Look-for Protocol

IUS

The Revised Standards for Mathematical Practice (SMPs) Look-for Protocol measures observable indicators related to K–12 teachers' promotion of the SMPs (Bostic & Matney, 2016; Bostic, Matney, & Sondergeld, 2019; E1). The SMPs are descriptions of mathematical behaviors and habits that teachers should promote and students should demonstrate (NGO Center & CCSSO, 2010; E2). It is a two-page form. The target population is K–12 mathematics teachers, inclusive of general and special education faculty (E4). Measurement contexts include both live and videorecorded lessons (E5). The Protocol is a single-form instrument with eight domains, one for each SMP, and three or four indicators for each SMP (E3). Presence of an indicator is marked dichotomously; no distinction is made about the quality of evidence related to an indicator. A total raw score may be summed across all indicators (e.g., 8/27; E7). Observers should not expect greater than 12 indicators during any single lesson. End users are responsible for calculating raw scores (E7). No cost is associated with use of the Protocol (E6). The Protocol was designed as a formative assessment and professional development evaluation tool to express the degree to which teachers promoted SMPs during an observed lesson (E8). Scores may be compared with means reported in Bostic et al. (2017). Observers using it as a formative assessment are strongly encouraged to debrief with teachers after the lesson as a coaching tool. Using it as a tool within program evaluation is appropriate when users want to explore changes in teachers' SMPs promotion, which may be influenced by an intervention (E9). Using it for evaluative purposes that can have professional or financial implications (i.e., high-stakes contexts) is not appropriate (E9). It is strongly recommended for use when an intervention's aim(s) closely align with the SMPs. Protocol users are expected to have a reasonable understanding of the SMPs that extends beyond a mere reading of them; a user should be able to describe a classroom-based scenario descriptive of each indicator. Finally, the Protocol is not appropriate for contexts in which SMPs are not a focus or applicable (E10).

Commentary

The second author served as lead developer for the Protocol and conducted nearly 1,000 hr of professional development focusing on K–12 instruction that promotes the SMPs. Its need arose from a lack of instruments that explicitly examined how teachers promoted the SMPs and from a desire to capture changes in teachers' instruction. Score interpretations provide users with information about teachers' instruction within a single instance and may be used in conjunction with other instruments to construct a profile of teachers' instruction. Knowledge of the Protocol's IUS supplies users with information about its scope and cautions against potential misuses. That is, it is intended for research purposes, evaluation of professional development initiatives related to the SMPs, and coaching; it is not an instrument to make high-stakes decisions. Hence, articulating the IUS for the Protocol communicates that its results capture a moment in time and cannot necessarily describe teaching outside of observed instruction. The Protocol started with an initial version, constructed by Fennell et al. (2013), that lacked much validity evidence and was created as a coaching tool. Those authors fully supported the validation process as expert panel members, and thus this instrument acts as an example of drawing from existing scholarship and working with original authors with a goal of creating tools with robust validity arguments.

Conclusion

With this Research Commentary, we hope to stimulate and elevate conversation within the field of mathematics education around argument-based validation, interpretation and use of scores, and assisting end users in appropriate identification and use of instruments. We make two recommendations to the mathematics education research community.

In a similar vein as Sztajn's (2011) recommendation for standards for reporting on professional development research, we recommend that all instrument developers supply an IUS. We offer the recommended IUS elements in this Research Commentary as a starting place for these discussions. Ideally, these recommendations will be iteratively improved as others reflect on them through discussion and use in practice. Through this process, the field can develop a shared understanding of the elements that are critical to examine and communicate to others related to instrument development, selection, and use, and through this process improve mathematics education research that makes use of instrument scores.

We recognize that more is needed in addition to the IUS to improve instrument usage in mathematics education research. We also recommend the development of instrument repositories as a tool for improving access to research that makes use of instruments. The IUS elements we have proposed could be used as a structure within the repository to communicate essential elements to end users. Repositories would reduce the burden on end users to find instruments and, by using the IUS elements as a structure, could encourage developers to focus on these elements, ultimately improving the practice of measurement in mathematics education research.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/uploads/7/6/6/4/76643089/ standards_2014edition.pdf

Barker, B. S., & Ansorge, J. (2007). Robotics as means to increase achievement scores in an informal learning environment. *Journal of Research on Technology in Education*, *39*(3), 229–243. https://doi.org/10.1080/15391523.2007.10782481

Bostic, J. D., Krupa, E. E., Carney, M. B., & Shih, J. C. (2019). Reflecting on the past and looking ahead at opportunities in quantitative measurement of K–12 students' content knowledge. In J. D. Bostic, E. E. Krupa, & J. C. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 205–229). Routledge. https://doi.org/10.4324/9780429486197-9

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education*, *24*(1), 5–31. https://doi.org/10.1007/s10857-019-09445-0

Bostic, J. D., & Matney, G. (2016). Leveraging modeling with mathematics-focused instruction to promote other Standards for Mathematical Practice. *Journal of Mathematics Education Leadership*, *17*(2), 21–33.

Bostic, J., Matney, G., & Sondergeld, T. (2017). (Re)considering teachers' promotion of the Standards for Mathematical Practice. In T. A. Olson & L. Venenciano (Eds.), *Proceedings for the 44th annual meeting of the Research Council on Mathematics Learning* (pp. 1–8). Research Council on Mathematics Learning.

Bostic, J. D., Matney, G. T., & Sondergeld, T. A. (2019). A validation process for observation protocols: Using the *Revised SMPs Look-for Protocol* as a lens on teachers' promotion of the standards. *Investigations in Mathematics Learning*, *11*(1), 69–82. https://doi.org/10.1080/19477503.2017.1379894

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2017). *The twentieth mental measurements yearbook*. University of Nebraska Press.

Carney, M., Crawford, A., Siebert, C., Osguthorpe, R., & Thiede, K. (2019). Comparison of two approaches to interpretive use arguments. *Applied Measurement in Education*, *32*(1), 10–22. https://doi.org/10.1080/08957347.2018.1544138

Carney, M. B., & Smith, E. (2017, September 12–14). *Using instrument development processes to iteratively improve construct maps: An example in proportional reasoning* [Paper presentation]. National Council on Measurement in Education first special conference on classroom assessment, Lawrence, KS, United States.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70*(5), 732–743. https://doi.org/10.1177/0013164410379323

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*(3), 397–412. https://doi.org/10.1177/0013164407310130

Fennell, F., Wray, J., & Kobett, B. M. (2013, January 24–26). *Using look for's to consider the Common Core content standards and Standards for Mathematical Practice* [Paper presentation]. Seventeenth annual meeting of the Association of Mathematics Teacher Educators, Orlando, FL, United States.

Hill, H. C., & Shih, J. C. (2009). Examining the quality of statistical mathematics education research. *Journal for Research in Mathematics Education*, *40*(3), 241–250. https://doi.org/10.5951/jresematheduc.40.3.0241

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Lobato, J., & Ellis, A. B. (2010). *Developing essential understanding of ratios, proportions, and proportional reasoning for teaching mathematics: Grades 6–8*. National Council of Teachers of Mathematics.

Minner, D., Erickson, E., Wu, S., & Martinez, A. (2013). *Compendium of research instruments for STEM education: Part 2. Measuring students' content knowledge, reasoning skills, and psychological attributes*. Community for Advancing Discovery Research in Education. http://www.cadrek12.org/sites/default/files/Student%20Compendium%20of%20STEM%20instruments%20Part%202%20with%20Addendum_May%202013.pdf

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. http://www.corestandards.org.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME *Standards for Educational and Psychological Testing*? *Educational Measurement: Issues and Practice*, *33*(4), 4–12. https://doi.org/10.1111/emip.12045

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, *31*(2), 165–174. https://doi.org/10.1080/08957347.2017.1408628

Sztajn, P. (2011). Standards for reporting mathematics professional development in research studies. *Journal for Research in Mathematics Education*, *42*(3), 220–236. https://doi.org/10.5951/jresematheduc.42.3.0220

Validity Evidence for Measurement in Mathematics Education. (n.d.). *February 2020 Las Vegas V-M^{2}ED Conference*. https://sites.ced.ncsu.edu/mathedmeasures/conference/february-2020-las-vegas-conference/

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 117–132. https://doi.org/10.1080/09695941003693856

Authors

Michele B. Carney, Department of Curriculum, Instruction, and Foundational Studies, Boise State University, 1910 University Dr., Boise, ID 83725; michelecarney@boisestate.edu

Jonathan Bostic, School of Teaching and Learning, Bowling Green State University, 529 Education Building, Bowling Green, OH 43403; bosticj@bgsu.edu

Erin Krupa, Department of STEM Education, North Carolina State University, Campus Box 7801, Raleigh, NC 27695; eekrupa@ncsu.edu

Jeff Shih, Department of Teaching and Learning, University of Nevada–Las Vegas, CEB 358, Las Vegas, NV 89154; jshih@unlv.nevada.edu

Submitted November 20, 2020
Accepted March 29, 2021
doi:10.5951/jresematheduc-2020-0087