ABSTRACTS FOR ASSESSMENTS: DESCRIBING A SUMMARY STATEMENT

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

Erin Krupa
North Carolina State University
eekrupa@ncsu.edu

Quantitative assessment development is a challenging process. The ways in which an assessment might be used, as well as how its score can be interpreted should be clear to intended users. This manuscript provides a discussion about important and useful elements that should be provided by assessment developers. In turn, this information can foster greater usability and portability of quantitative assessments, which can support scholarship focusing on a specific issue.

Keywords: Assessment

Quantitative research requires the use of measures, instruments, or assessments that allow users to draw conclusions from data they collect. When a research purpose aligns with a previously created assessment, then it seems plausible to use it. On the other hand, if there are no assessments to measure a desired construct, then an individual or team must choose whether to develop one. In either case, a statement describing how to use the assessment might be employed to make scholarly decisions. At present, there is little guidance about what goes into such a statement for potential users of quantitative assessments. The purpose of this manuscript is to present list of recommendations to include in a summary statement for an assessment and then provide an example summary statement to highlight the recommendations.

Related Literature

The *Standards for Educational and Psychological Testing* ([*Standards*] AERA et al., 2014) describe five validity sources: test content, response process, relations to other variables, internal structure, and consequences from testing/bias. Reliability is a related component of the *Standards* but is not one of the five sources. These sources describe categories in which evidence may be grouped in order to make score interpretations and effectively use a measure. The *Standards* note that it is inappropriate to use phrases such as "the validity of the test" or that a test is valid and instead, encourage a focus on validation as "...the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (p. 11). Thus, validation is a process and the validity evidence supports or refutes the score interpretations and uses (Kane, 2006; 2016). This may seem like a small language shift but such a shift has serious implications because a valid test conveys a different idea as opposed to a valid score interpretation from a test.

Historically speaking, this shift has been slow to happen in some mathematics education communities. For example, Bostic and colleagues (2019) analyzed *Journal for Research in Mathematics Education* (1970-2017) manuscripts examining elementary students' learning outcomes to discern (a) whether validity evidence for uses was provided and if so, (b) how was the evidence presented. One result from that analysis was that seven of 97 manuscripts (7%) that used quantitative measurement with elementary students' outcomes described any validity evidence associated with their instruments. It was most common to describe test content evidence from an expert panel review as well as a reliability statistic. From the 1980s onward, it became increasingly common for *Journal for Research in Mathematics Education* authors to discuss an author-created measure. It is difficult to determine whether the instruments described

Olanoff, D., Johnson, K., & Spitzer, S. (2021). *Proceedings of the forty-third annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Philadelphia, PA.

in these published articles might be useful for a scholar's purpose because it is unclear how to administer, score, and interpret results from the measure. A framework to help measure developers describe these aspects has potential to increase assessment usability and assist scholars seeking to conduct quantitative research within mathematics education contexts. While in-depth descriptions of assessments are still warranted (e.g., validation research), succinct and explicit summaries are needed for readers to quickly and effectively discern whether to consider an assessment for a desired use. Concomitantly, peer-reviewed assessment and validity research within mathematics education contexts has increased substantially in the last five years; hence, a need to have a shared framework for communicating a summary statement related to a quantitative assessment. This manuscript responds to the question: What should a summary statement for an assessment contain?

Method

Context and Participants

This qualitative study stems from work during a NSF-sponsored conference. The conference lasted two days. Attendees were selected from an application process that brought together 41 scholars with expertise in mathematics education, mathematics, psychometrics, or applied measurement. This group included 35 terminally-degreed individuals working in industry and university settings, as well as six graduate students. A major goal of the conference was to identify a set of recommendations for the elements to include in an interpretation and use statement (aka purpose statement).

Data Collection and Analytical Process

A set of elements were initially generated by the conference leaders. These elements were based on important elements highlighted in the *Standards* and provided to conference participants. Conference attendees were asked to draft an example summary statement for an assessment around a construct of their choice using these elements as a starting point. They were asked to note elements to include and eliminate from the provided list, and to add additional elements to discuss for inclusion. This small group work time was followed by a whole group discussion on the common elements to include in the summary statement. 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. A revised document was used by small groups of participants to draft a new summary statement and provide feedback 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, videorecordings from the conference, and field notes from the whole group discussion were analyzed following the conference and used to craft a set of reporting recommendations for elements of the summary statement. Four researchers (i.e., the leaders of the conference) used inductive analysis (Creswell, 2012) as a tool to develop the summary statement. The inductive analysis started with re-reading (or re-listening) to materials (e.g., written work and recorded statements from the conference). Step two was to make memos consisting of initial ideas stemming from this examination of the data. Step three was to reflect on those memos as a way to synthesize them into support (or not) for aspects of the summary statement. This is needed as evidence to ground the summary statement in validity. Step four was to search for evidence within the data sets to support components of the summary statement. Step five was to search the data for counter evidence. Impressions with a paucity of counter evidence and a large set of evidence were

retained. The sixth and final step was crafting clearly elements to share broadly as a summary statement.

Findings: A summary statement

We present the recommendations for a summary statement first, then describe some of the comments surrounding its development. The ten elements were grouped to better visualize three different aspects of a quantitative measure: Construct articulation, operationalization and administration, and scores. Construct articulation provides justification for measuring the construct and clarifying its importance. Operationalizing and administering the measure is intended to give information about how the measure should be used. Scores and scoring describe aspects related to scoring and the limitations/delimitations related to the measures' scores.

Aspect	Interpretation and Use Element
Construct Articulation	#1. Why measure this construct?
	#2. Why is it important to measure this construct?
Operationalization and Administration	#3. How is the construct measured?
	#4. Who is the target population?
	#5. What is the intended context for administration?
	#6. What are associated costs with using the instrument?
Scores and Scoring	#7. How are scores determined?
	#8. What are intended interpretations for scores?
	#9. How should scores be used?
	#10. What known warnings or cautions are important to consider?

Figure 1. Summary statement to describe score interpretations and uses of a quantitative assessment

There was consensus that the summary statement should be written for an end-user to make a decision about (a) whether the score interpretation from an assessment aligns with an intended use and (b) the degree to which the assessment aligns with a desired purpose. One of the participants working in the assessment industry, Melissa, said that "You still need the details for an instrument and its uses. A summary statement is a quick read." Lucas, a university faculty, said that "This summary statement can tell you whether the instrument actually measures what it says it does. It can also show where there are gaps in the validity argument to further explore." As a result of video, audio, and written data, we reached the conclusion that the summary statement provided necessary and sufficient evidence for potential end-users.

An Instantiation of the Summary Statement

We present an example summary statement for a problem-solving measure developed by an author of this manuscript. There are numerous peer-reviewed manuscripts detailing validity

evidence and arguments for this problem-solving measure (PSM); hence, it provides a brief overview for potential measure users and administrators. It should be interpreted cautiously and provide readers with an example of a potential summary statement for an actual instrument.

The PSM3 measures students' problem-solving performance within the context of the third-grade Common Core State Standards for Mathematics (CCSSO, 2011). Past research has demonstrated that problem-solving measures (a) are large-scale in nature (e.g., PISA), (b) measure problem solving but the mathematics content does not align with instructional standards, or (c) measure problem solving without drawing on mathematics content (see Bostic & Sondergeld, 2015). Thus, the PSM3 fills a need as a problem-solving measure that aligns with instructional standards used in many states within the USA. It has 15 items displayed as word problems. Each is presented as a constructed response task. Students are asked to clearly write their answer on a provided line. The target population is English-speaking, grade-level students. PSM3 administration is typically performed during instruction for and can last 120 minutes; however, most students finish within 90 minutes. There is no difference in students' outcomes due to the completing the PSM in one sitting or across multiple sittings (e.g., six, 20-minute sittings). Calculators are not allowed for administration. Those interested in using the PSMs may contact the authors for pricing. Each item is scored dichotomously, which conveys the same information as partial credit scoring (Carney et al., accepted). Respondents' scores may be calculated as percent correct. Scores may also be analyzed using Rasch to explore how students' performance compares to norms. Results from Rasch analysis should be interpreted as information about students' problem-solving performance related to CCSSM content. Such Rasch results also convey students' outcomes compared to peers and norms. PSMs are designed to complement other data about students' mathematics outcomes and be interpreted as a single touchpoint of students' outcomes. PSM data are suitable for research, evaluation, and school-based needs and as seen in this manuscript, robustly address validity *Standards* (AERA et al., 2014). Results are not intended to track students into different mathematics classes.

Discussion and Implications

This summary statement is intended to provide scholars working within mathematics education contexts a shared perspective to convey information about their quantitative assessments. It functions much like an abstract serves a manuscript or proposal – offering at-a-glance information. This summary statement also addresses the five validity sources, which may be further unpacked. For example, the ways in which scores are analyzed using Rasch analysis tells a reader that the PSM3 results are measured in logit units, which cues a reader to deciding whether that suits their needs. One implication from this research is to further engage the mathematics education scholarly community in ways that encourage sharing measures, replication studies, and offer greater access to quantitative measures. Kane (2016) and the Standards (AERA et al., 2014) have recommended that clearly identifying key information about measures has potential to improve measurement practices.

Acknowledgments

This work was supported by grants 1720646, 1720661, 1920619 and 1920621 from the National Science Foundation. Further, the authors want to acknowledge Michele Carney, Jeffrey Shih, and the V-M^{2}ED community for their efforts and support of this work. Any opinions expressed in this manuscript are those of the authors and do not 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.* Washington, DC: American Educational Research Association.

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*.

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.

Carney, M., Bostic, J., Krupa, E., & Shih, J. (accepted). Instruments and use statements for instruments in mathematics education. *Journal for Research in Mathematics Education*. Accepted for publication. Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics*. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Creswell, J. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Kane, M. T. (2006). *Validation*. In R. L. Brennan, National Council on Measurement in Education & American Council on Education (Eds.), Educational measurement (pp. 17-64). Westport, CT: Praeger Publishers.

Kane, M. T. (2016). *Validation Strategies: Delineating and Validating Proposed Interpretations and Uses of Test Scores. I*n S. Lane, M. Raymond & T. M. Haladyna (Eds.), Handbook of Test Development (Vol. 2nd; pp. 64-80). New York, NY: Routledge.