MEASURING WHAT WE INTEND: A VALIDATION ARGUMENT FOR THE GRADE 5 PROBLEM-SOLVING MEASURE (PSM5)

Jonathan Bostic
Bowling Green State University
bosticj@bgsu.edu

Gabriel Matney
Bowling Green State University
gmatney@bgsu.edu

Toni Sondergeld
Drexel University
tas365@drexel.edu

Gregory Stone
Metriks Amerique
gregorystone@metriks.com

Abstract

The purpose of this proceeding is to share a validity argument for the Problem-solving Measure for grade 5 (PSM5). The PSM5 is one test in the PSM series, which is designed for grades 3-8. PSMs are intended to measure students' problem-solving performance related to the Common Core State Standards for Mathematics (i.e., content and practices). In addition to sharing validity evidence connected to the PSM5, we discuss implications for its use in current research and practice.

Acknowledgement

Ideas in this manuscript stem from grant-funded research by the National Science Foundation (NSF #1720646, 1720661). Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation.

Introduction

Problem solving is found in both the Standards for Mathematics Content and Standards for Mathematical Practice (Common Core State Standards Initiative [CCSSI], 2010). There is no doubt about its importance as part of classroom instruction (National Council of Teachers of Mathematics, 2000). Because it is an important part of instruction, it should be assessed in a way that provides students, teachers, and other school personnel with valuable information. Unfortunately, there continues to be few quantitative measures of problem solving that align with mathematics standards (Bostic, Krupa, & Shih, 2019; Bostic, Sondergeld, Folger, & Kruse, 2017). The purpose of this manuscript is to provide a validation argument for a new test within a series of Problem-solving Measures (PSMs). The PSMs are designed for students learning mathematics designed for grades 6, 7, and 8. The test in the present study is meant for grade 5 students; hence, it is called the PSM5.

Relevant Literature

Problems and Problem Solving

There are entwined, mutually beneficial frameworks intended to frame the purpose and intent of the PSM5 and its items, specifically problem solving and problems. First, problems were

defined using two frameworks. The first framework was Schoenfeld's (2011) notion that problems are tasks for a problem solver such that (a) it is unclear whether there is a solution, (b) it is unknown how many solutions exist, and (c) the pathway to the solution is unclear. The second framework for problems stems from work conducted by Verschaffel and colleagues (1999). Problems are (a) open, (b) complex, and (c) realistic tasks for an individual. Open tasks can be solved using multiple developmentally-appropriate strategies. Complex tasks are not readily solvable by a respondent and require productive thinking. Realistic tasks may draw upon real-life experiences, experiential knowledge, and/or believable events. Problems are quite different from exercises, which are meant to support an individual's efficiency with a known procedure (Kilpatrick, Swafford, & Findell, 2001). These two frameworks for problems are synergistic and provided PSM5 developers a roadmap for what should be included in tasks.

The framework for problem solving that guides PSM development is a process of "several iterative cycles of expressing, testing and revising mathematical interpretations – and of sorting out, integrating, modifying, revising, or refining clusters of mathematical concepts from various topics within and beyond mathematics" (Lesh & Zawojewski, 2007, p. 782). Such a problem-solving perspective requires tasks that encourage students to engage in productive, reflective, goal-oriented problem solving (Schoenfeld, 2011; Yee & Bostic, 2014). Problem solving takes substantially more cognitive effort compared to executing procedures to complete exercises (Polya 1945/2004).

Validity and Validity Arguments

Validation is an important part of the assessment development process and while it, "may not be easy…it is generally possible to do a reasonably good job of [it] with a manageable level of effort" (Kane, 2016, p. 79). Validation, broadly speaking, involves the process of gathering evidence and constructing an argument that connects an instrument's outcomes and/or interpretations from it to its designed purpose (Kane, 2006; 2012). Validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014, p. 11). Second, this research draws upon the *Standards* (AERA et al., 2014), which describe five sources of validity as necessary facets for assessment development: test content, response process, internal structure, relations to other variables, and consequences from testing. Third, a validation argument

typically follows a specific format (e.g., Kane, 2016; Pellegrino, Dibello, & Goldman, 2016; Wilson & Wilmot, 2019) to convey validity evidence. For this manuscript, we use *argument* to indicate a "coherent series of reasons, statements, or facts intended to establish a point of view" (Merriam-Webster, 2018). Therefore, a validation argument serves to inform readers of the validity evidence and why it justifiably grounds the implications and results from an instrument. To that end, the research question for the present study was: What is the validity argument for the PSM5?

Method

This study draws upon a design science approach (Middleton, Gorard, Taylor, & Bannan-Ritland, 2003) and connects with recent literature that validation is a methodology within mathematics education research (Jacobsen & Borowski, 2019). Design science research is valuable for creating products that can be evaluated, refined, and re-evaluated. Jacobsen and Borowski argued that validation work serves as a methodology unto itself because there are specific characteristics of such work. For the purposes of this study, the *Standards* (AERA et al., 2014) were chosen as a mechanism to convey the validity argument for this manuscript. This approach for the validity argument was used for previous research examining the PSMs.

The *Standards* (AERA et al., 2014) advocate for assessment developers to gather evidence for the five sources; however, the quality of evidence rather than the quantity of evidence is more important. Large amounts of evidence for two sources is not sufficient though (Bostic, 2017) for a validity argument. Past research that has drawn solely upon test content and internal consistency evidence does not provide a sufficiently robust validity argument such that others might trust that the results and interpretations are valid (Bostic, 2017).

Instrument and Participants

There were two groups of participants involved in this study. All names are pseudonyms and the study was approved by the Institutional Review Board. The first group was fifth-grade students. Fifth-grade students participated in think-aloud interviews, consequences from testing/bias interviews, and actual testing of the PSM5. Students were purposefully selected from rural, suburban, and urban districts within the Midwest USA. Seventy-three students in total participated in think alouds and 335 students participated in PSM5 test administration. The second group of participants were fifth-grade teachers, mathematics teacher educators whose focus is elementary grade levels, and mathematicians who have expertise is teaching

mathematics content for elementary teachers. All adult participants for the expert panel communicated having sufficient understanding of the Common Core State Standards for Mathematics and agreed to review the PSM5 for content and potential bias.

The PSM5 that students completed contained 18 items meant to measure students' problem-solving performance within the context of Common Core State Standards for Mathematics Content (SMC) and Practices (SMPs). There are at least three items for each of the five mathematical domains found in the fifth-grade SMCs (i.e., Operations and Algebraic Thinking, Number and Base Ten, Number and Fractions, Geometry, and Measurement and Data). A sample PSM5 item reads: "The State Nut Company buys 22 pounds of pecans, 30 pounds of walnuts, 30 pounds of peanuts, 25 pounds hazelnuts, and 30 pounds of almonds. They sell mixed-nuts in 2.5-pound containers, which contain exactly 0.5 pounds of each nut type. How many containers will they make?". Items have been previously reviewed by an expert panel and those results were reported in Bostic, Matney, Sondergeld, & Stone (2018).

Data Collection and Analysis

Table 1 provides an outline of data collected, analysis technique used, and how it connects to the validity evidence framework. Expert panel reports were gathered from multiple fifth-grade mathematics teachers who had more than three years teaching experience in that grade, mathematics teacher educators, and mathematicians. Their reports provided feedback on connections to mathematics content, mathematics practices (CCSSI, 2010), and potential areas of bias. Think alouds were conducted with fifth-grade students several months prior to test administration and immediately following test administration. The goals for early think alouds were to explore ways that students might respond to PSM5 items. Think alouds following test administration were conducted to discern students' feelings and affect after testing. These qualitative data were analyzed using thematic analysis, similar to past PSM analyses (see Bostic & Sondergeld, 2015; Bostic, Sondergeld, Folger, & Kruse, 2017). Quantitative data collection for relations to other variable evidence included collecting demographic evidence about the 335 respondents. Students' responses to the items were analyzed using Rasch modeling to interpret students' and items' qualities. Finally, bias was investigated using independent samples t-tests and Rasch (Rasch, 1960/1980) techniques to explore whether there were any differences in students' performance.

Table 1. *Connections between validity evidence, data collection, and data analysis*

Validity Evidence Source	Data collected	Data analysis technique
Test Content	Expert panel reports (qualitative)	Thematic analysis (Creswell, 2012; Hatch, 2002)
Response processes	Think-aloud data with representative purposeful sampling of students (i.e., different ability levels, genders, and geographic context) (n=73; qualitative)	Thematic analysis (Creswell, 2012; Hatch, 2002)
Relations to other variables	Ability level, gender, and geographic contexts (quantitative)	Independent samples t-tests
Internal Structure	Test results from 335 respondents across 4 schools (quantitative)	Rasch modeling
Consequences from testing/bias	Expert panel reports, think-alouds with purposeful, representative sample of students following test administration, teacher interviews following test administration, and analyzing relations to other variables evidence (mixed methods)	Thematic analysis (Creswell, 2012; Hatch, 2002) Independent samples t-tests

Results

The results from validity evidence analysis are presented in relation to the five sources. First, the experts provided positive feedback indicating that the PSM5 items were connected to fifth-grade SMCs, address the SMPs, could be solved using multiple developmentally-appropriate strategies, were complex enough to be considered problems, and drew upon realistic contexts. Second, response processes results indicated that students were able to use appropriate mathematical strategies while problem solving PSM5 items. Readability of the items was not an issue, as evidence by students' abilities to read and understand what each question asked. Third, evidence about relations to other variables suggested that the PSM5 functioned as desired. Independent samples t-tests comparing ability levels, gender, and ethnicity all reported expected results. Higher ability students outperformed average-ability and below average-ability students. There were no statistically significant differences between white and non-white students as well as no differences between performances by gender. There were also no statistically significant differences between students from different geographic locations (i.e., rural, suburban, and urban). Some items indicated that females performed better than males whereas other items suggested that males performed better than females, which is normal for an entire measure.

Collectively speaking however, there was no overall difference between male and female performance on the PSM5. Fourth, internal structure evidence was evident that psychometrically the test functioned effectively. Separation and reliability scores of 2.00 and .80 are considered good while 3.00 and .90 are considered excellent (Duncan, Bode, Lai, & Perera, 2003). Person separation (i.e., number of distinct groups that can be classified on the variable) and reliability were trending towards good (i.e., 1.6 and .73 respectively). Item separation and reliability exceeded the threshold for excellent (7.0 and .98 respectively). Finally, the expert panel and students reported that they did not experience or notice any bias in the PSM5. Post-test administration interviews revealed that students felt that the test was similar to a unit test. Students reported feeling satisfied that their results might be used to inform teachers' instruction. Bias analyses from quantitative data revealed that across the test as a whole, bias was not weighted towards one group (e.g., males or females).

Discussion and Next Steps

Taken collectively, the validity evidence indicated that the PSM5 functions as intended. This evidence parallels the quality of validity evidence seen in the PSM6-8 series, which addresses expectations described in the *Standards* (AERA et al., 2014). This new PSM5 also extends the PSM series (see Bostic & Sondergeld, 2015; Bostic et al., 2017) into elementary grade levels. Work on the PSM3 and PSM4 is running in parallel to the PSM5, which will offer an assessment series that has potential to examine students' progress from elementary school into middle school mathematics content.

Drawing upon the design-science approach to this work, the development team has revised the PSM5 with the intent to improve the person separation values and to shorten the test. Both features are likely to improve quality and result in better psychometric values. While person separation and reliability are lower than desired, measuring students' problem solving can present issues because problem solving is more difficult than performance on exercises or other routine mathematics items (Bostic & Sondergeld, 2015). Thus, it might be expected to have low person separation scores. Another next step is revising the PSM5 to include fewer items; however, drawing upon high quality items. The results for this validation study stem from data collected May 2019. A revised PSM5 was piloted during the fall of 2019, which will generate new internal structure findings to report. However, it is unlikely to change validity evidence for the other sources.

Final Thoughts

There are few quantitative instruments that adhere to the *Standards* (Krupa, Bostic, & Shih, 2019). Too often, research has drawn upon poorly constructed measures that lack validity evidence to justify the results and implications from the measures (Krupa et al., 2019). This validation study provides evidence that indicates how results and interpretations from the PSM5 are connected to the actual measure. The PSM5 offers teachers, researchers, and school districts a means to effectively capture students' learning and use these findings to make instructional decisions.

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.

Argument. (2018). *Merriam-webster.com.* Retrieved from ttps://www.merriam-webster.com/dictionary/argument

Bostic, J., Krupa, E., & Shih, J. (2019). Introduction: Aims and scope for *Assessment in mathematics education contexts: Theoretical frameworks and new directions*. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 1-11). 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*.

Bostic, J., Sondergeld, T., Folger, T. & Kruse, L. (2017). PSM7 and PSM8: Validating two problem-solving measures. *Journal of Applied Measurement, 18*(2), 151-162.

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. (2018). Improving test development reporting practices. In L. Venenciano & A. Sanogo (Eds.), *Proceedings of the 45th Annual Meeting of the Research Council on Mathematics Learning* (pp. 57-64). Baton Rouge, LA.

Bostic, J., Matney, G., Sondergeld, T., & Stone, G. (2018). Content validity evidence for new problem-solving measures (PSM3, PSM4, and PSM5). In T. Hodges, G. Roy, & A. Tyminski (Eds.), Proceedings for the 40h Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (pp. 1641). Greenville, SC.

Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics*. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

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

Duncan, P., Bode, R., Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84*, 950-963.

Hatch. J.A. (2002). *Doing qualitative research in educational settings.* Albany, NY: State University of New York Press.

Jacobsen, E. & Borowski, R. (2019). Measure Validation as a Research Methodology for Mathematics Education. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge (pp. 40-62)*. New York, NY: Routledge.

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

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17.

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

Kilpatrick, J., Swafford, J., and Findell, B. (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.

Krupa, E., Bostic, J., & Shih, J. (2019). Validation in mathematics education: An introduction to *Quantitative Measures of Mathematical Knowledge: Researching Instruments and Perspectives.* In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative Measures of Mathematical Knowledge: Researching Instruments and Perspectives* (pp. 1-13). New York, NY: Routledge.

Lesh, R., & Zawojewski, J. (2007). Problem solving and modeling. In F.K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning: A project of the National Council of Teachers of Mathematics*. (pp. 763-803). Charlotte, NC: Information Age.

Middleton, J., Gorard, S., Taylor, C., & Bannan-Ritland, B. (2008). The "compleat" design experiment. In A. Kelly, R., Lesh, & J. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics teaching and learning* (pp 21-46). New York, NY: Routledge.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Pellegrino, J., DiBello, L., & Goldman, S. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.

Polya, G. (1945/2004). *How to Solve It.* Princeton, NJ: Princeton University Press.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Copenhagen: Denmarks Paedagoiske Institut.

Schoenfeld, A. H. (2011). How we think: A theory of goal-oriented decision making and its educational applications. New York, NY: Routledge.

Verschaffel, L., De Corte, E., Lasure, S., Van Vaerenbergh, G., Bogaerts, H., & Ratinckx, E. (1999). Learning to solve mathematical application problems: A design experiment with fifth graders. Mathematical Thinking and Learning, 1, 195-229.

Wilson, M., & Wilmot, D.B. (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)*. New York, NY: Routledge.

Yee, S., & Bostic, J. (2014). Developing a contextualization of students' mathematical problem solving. *Journal for Mathematical Behavior*, *36,* 1-19.