# THREE STEPS ForWARD: VALIDITY EVIDENCE FOR THE PSM3

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

Toni A. May
Drexel University
Tas365@drexel.edu

Gabriel Matney
Bowling Green State University
gmatney@bgsu.edu

Gregory Stone
MetriKs Amerique
gregory@metriks.com

*This paper's purpose is to discuss validity evidence related to a third-grade problem-solving measure (PSM3). PSM3 is connected to a series of tests designed to measure students' problem-solving performance aligned with the Common Core State Standards for Mathematics. Multiple validity sources are drawn together to support the PSM3's interpretations and uses.*

Keywords: Assessment; Elementary School Education, Problem Solving

Problem solving is central to mathematical work (National Council of Teachers of Mathematics [NCTM], 2000, 2014) and is a core part of the Common Core State Standards, which were adopted by 42 of 50 states (Common Core State Standards Initiative [CCSSI], 2010). Problem solving is found in every grade-level across the Standards for Mathematics Content (SMCs) and is described in the first Standard for Mathematical Practice (SMP; CCSSI, 2010). The notions of problem and problem solving are pervasive across the Standards for Mathematical Practice (e.g., "Make sense of problems and persevere in solving them, CCSSI, 2010, p. 6) as well as the Standards for Mathematics Content (e.g., "Solve two-step word problems using the four operations", CCSSI, 2010, p. 23) and therefore should be a part of mathematics assessments. Bostic and colleagues (2015; 2017) reported that problem-solving tests used in scholarly studies tend to fall into three categories: large-scale assessments, measures of mathematical problem-solving distinct from curricular standards, and problem-solving assessments focusing on nonmathematical elements. Unfortunately, few mathematical quantitative instruments used with elementary students have reported validity evidence supporting their uses (Bostic et al., 2019). This study fills a gap in the literature by providing validity evidence for a problem-solving measure connected to curricular standards within elementary settings.

## Related Literature

Multiple definitions and frames for mathematical problem solving exist. This study is guided by Lesh and Zawojewski's (2007) modeling-influenced perspective on problem solving: "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" (p. 782). Such a problem-solving perspective requires tasks that encourage students to engage in productive, reflective, goal-oriented problem solving. While there are multiple frames and definitions for what counts as a problem, this study draws upon Schoenfeld's (2011) features of a problem: (a) it is unknown whether a solution exists, (b) a solution pathway is not readily determined, and (c) there exists more than one way to answer the task. Problem solving happens when a task is a problem, not an exercise, for an individual (Polya, 1945/2004; Schoenfeld, 2011); hence a key component to problem solving is a problem.

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.

The PSMs contain word problems and were designed using Verschaffel et al.'s (1999) characterization of word problems: *Open* word problems can be solved in different ways and offer learners multiple entry points. *Realistic* word problems draw on a problem solver's experiential knowledge and engage the student in a real-world task. *Complex* word problems require an individual to employ sustained reasoning. Communicating definitions is important to this study because developing summary (aka purpose) statements within validation work is derived from purposeful choices and in turn, informs users what the instrument can and cannot do (Carney et al., accepted). These statements are like an abstract for an assessment in that they convey essential information for potential measure users and administrators.

This study draws upon the Standards (AERA et al., 2014) to communicate evidence and connect it to interpretations and use. Aspects of a test's interpretation and use include articulating a construct, describing test administration, and scoring (Carney et al., accepted). The research question for this study is: What validity evidence exists for the PSM3? This examination of the PSM3 builds upon work on past PSMs for grades 4-8 (see Bostic et al., 2015; 2017; 2020).

## Method

A design science framework (see Middleton, et al., 2008) guides this study to explore five sources of validity (see AERA et al., 2014): test content, response process, relations to other variables, internal structure, and consequences from testing. Only test content, response processes, and internal structure will be highlighted in this paper due to page limitations. Test content evidence provides a connection between content described in items on a test and the intended construct (AERA et al., 2014; Sireci & Faulkner-Bond, 2014). Reviews from an expert panel are a common and appropriate approach for discerning the degree to which there is a match (AERA et al., 2014). Response process evidence explores if respondents behave in ways that are intended or desirable (Padilla & Benitez, 2014). Think alouds are typical approaches to gather response process evidence for problem-solving tests (Leighton, 2017). Internal structure evidence suggests the degree to which items conform to a desired construct (AERA et al., 2014). Rasch techniques as well as classical test theory approaches are both adequate, yet each approach is beholden to differing assumptions (Rios & Wells, 2014). Qualitative data and analyses were used with test content and response process evidence. Quantitative data and analyses were employed to explore internal structure evidence.

### Measure

The PSM3 is composed of 15 word problems with three items coming from each of the five SMC content domains: Operations and Algebraic Thinking, Numbers in Base Ten, Number and Fractions, Measurement and Data analysis, and Geometry. A sample PSM3 item reads "Beth is coloring a picture using crayons. The box of crayons has 6 blue crayons, 4 yellow crayons, 8 green crayons, and 6 red crayons. What fraction of the box of crayons is green?" The PSM3 is designed to measure mathematical problem-solving in relation to third-grade mathematics standards.

### Data Collection

To address test content, expert panels were conducted with three grade-level mathematics teachers, two terminally-degreed mathematics educators with expertise in elementary mathematics (grades K-6), and one terminally-degree mathematician. Mathematics teachers were current grade three mathematics teachers who had at least four years teaching experience and at least two years teaching third grade. The mathematics educators have elementary teaching experience and have published and presented peer-reviewed work on elementary mathematics

teaching. The mathematicians has experience working with elementary teachers and communicated having read and discussed the Common Core State Standards with their university students. Mathematics teachers and teacher educators responded to the following questions: (1) Is the task a problem? (2) Is the task open? (3) Is the task realistic? (4) What Standard(s) for Mathematics Content are primarily addressed by this task? (5) What Standard(s) for Mathematical Content are primarily addressed by this task? The mathematician responded to questions #1-3 as well and additionally, (6) Describe the mathematics addressed by this task. What are two appropriate, grade-level problem-solving strategies? (7) Is the mathematics in the problem correct? (8) Is there a well-defined solution for the task? Items were reviewed once by the expert panel, revised, and then subjected to a second review when necessary. Each expert panel member submitted responses to these questions.

To address response processes, both 1-1 think alouds and whole-class think alouds were used. 1-1 think alouds were performed with a purposeful sample of 12 students consisting of varying mathematical abilities as report by their mathematics teachers (i.e., above average, average, and below average ability), male and female students, as well as white and non-white students. Ability-level judgments were gathered from teachers' views about students' classwork and prior assessment data. Whole-class think alouds (see Bostic et al., 2021) were conducted one year later with two unique sets of students ($n$=32). Think alouds were videotaped and student work was collected. Combining think-aloud formats allowed for greater and more diverse information about students' responses.

To address internal structure, third-grade students ($n$=290) across four Midwest districts completed the PSM3 in the last month of the academic year. Districts represent urban, suburban, and rural schools and each has unique populations consisting of different ethnic backgrounds, socio-economic status, and locations. Students with and without an identified disability completed the PSM3 per any Individualized Education Plan requirements. Based upon prior pilot administrations, teachers gave students approximately 90 minutes to answer the questions.

**Data analysis**

Expert panel reports and student think aloud data were analyzed using inductive analysis (Creswell, 2012) across three researchers, which maintains a parallel structure from previous peer-reviewed work (Bostic et al., 2015; 2017; 2020). The inductive analysis started with re-reading (or re-watching) to materials (e.g., written work and recorded statements from the conference). Next, we made memos consisting of initial ideas stemming from this examination of the data and later reflected on those memos to synthesize them into support (or not). Then, we sought evidence and counter evidence within the data sets to support our burgeoning themes. Impressions with a paucity of counter evidence and a large set of evidence were retained. Finally, we crafted a thematic statement representing the supporting data. Related to test content evidence, an intended goal was to discern the degree to which items were connected to the intended standards and addressed our selected framework for word problems. Related to response process evidence, an intended goal was to explore ways that students' responses aligned with our a priori conjectures in students' problem solving. Psychometric data analysis for internal structure used Rasch modeling (Rasch 1960/1980). PSM3 items were scored dichotomously by three scorers using a scoring key. Generally, it is important to look multiple components from Rasch analysis. First, separation and reliability values of 2.0 and 0.8 are considered good while 3.0 and 0.9 are excellent (Duncan et al., 2003). Rasch infit and outfit statistics (mean square values between 0.5 and 2.0) are considered acceptable and there should be no negative point-biserial statistics (Linacre, 2002).

## Findings

Themes for test content evidence were tasks were: complex enough to be considered problems for third-grade students, open, and solvable in multiple ways using grade-level strategies, and based upon realistic contexts that led to realistic solutions. Mathematicians confirmed three and sometimes four developmentally appropriate strategies that students might use to solve the word problems. Expert panel feedback consistently conveyed that tasks aligned with third grade content standards. One teacher shared a sentiment that others echoed: "These are appropriately difficult word problems that will make students think about the math they learn. These problems require more than just using a procedure." Finally, the expert panel conveyed that word problems met developmentally appropriate reading levels. A Flesch-Kincaid reading analysis confirmed (3.4 grade level). In sum, there was majority agreement between expert panel members and researchers' hypothesized content standards.

A theme about response process evidence was that students responded in anticipated ways. Average- and above-average performing students tended to provide more correct answers than below-average students. It was common for lower-performing students to combine numbers using symbolic notation without making sense of the quantities. In the crayon problem described earlier, there were many students who wrote a fraction that did not answer the question. When pressed to explain their thinking, we heard comments like Natasha's: "I made a fraction with the numbers like it says in the problem." All students were able to read the problems, which supported our finding that the PSM3 met grade-level reading expectations.

Psychometric findings support robust internal structure evidence. All items had acceptable infit (MNSQ Range 0.82-1.29) and outfit (MNSQ Range 0.68-2.00) measures, and no items had negative point biserial values. Rasch item reliability (0.93) and separation (6.43) were strong. Collectively, psychometric data suggest a unidimensional variable of problem solving has been created from items on the PSM3.

## Discussion & Limitations

The central aim for this study was to report test content, response process, and internal structure validity evidence for the *PSM3*. Synthesized findings suggest the validity evidence as being supportive of the following claims: (a) Mathematics content found on *PSM3* tasks addresses mathematics content described in grade-level standards; (b) Respondents solved *PSM3* tasks in anticipated ways; and (c) The *PSM3* appears to fit a unidimensional construct, which we characterize as mathematical problem solving. These findings connect back to three desired sources of validity: test content, response process, and internal structure (see AERA et al., 2014). Taken collectively, the *PSM3* is an instrument that may be useful for scholars interested in studying third-grade students' mathematical problem solving within instructional contexts using the Common Core State Standards for mathematics. Evidence for relations to other variables as well as consequences of testing/bias will be further investigated. The findings for this study are limited to native English speakers, which should be explored in subsequent studies.

## Acknowledgments

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.

## 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., 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., 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., Matney, G., Sondergeld, T., & Stone, G. (2020, March). *Measuring what we intend: A validation argument for the grade 5 problem-solving measure (PSM5). Validation: A Burgeoning Methodology for Mathematics Education Scholarship*. In J. Cribbs & H. Marchionda (Eds.), Proceedings of the 47th Annual Meeting of the Research Council on Mathematics Learning (pp. 59-66). Las Vegas, NV.

Bostic, J., Sondergeld, T, Matney, G., Stone, G., & Hicks, T. (2021). Gathering response process data for a problem-solving measure through whole-class think alouds. *Applied Measurement in Education, 34*(1), 46-60.

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.

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.

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

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.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16(*2), p. 878. Retrieved from https://www.rasch.org/rmt/rmt162f.htm

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.

National Council of Teachers of Mathematics (2014). Principles to Actions: Ensuring Mathematical Success for All. Reston, VA: Author.

Padilla, J-L., & Benitez, I. (2014). Validity evidence based on response processes. *Psichotherma, 26*, 136-144.

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.

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

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

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

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.