Title Page with Author Information

Authors: Jonathan David Bostic, Gabriel Matney, Timothy Folger, Noah Brown, Emily Evans, Toni Sondergeld, Gregory Stone

Proposal Type: Paper Session submission

Topic area: STEM Education

Abstract: Problem solving is a theme within mathematics instructional standards, which suggests that assessments must address mathematics content and practices. This study examines grades 3-5 students' strategy use on a test designed to measure students problem-solving performance within the context of mathematics standards used during instruction. The tests are called the Problem-Solving Measures. Results indicate that students use multiple and varied strategies to solve PSM items, which in turn supports prior validation studies.

Description of Presentation: First, we will describe the PSMs 3-5. Next, we will present sample items from the PSMs and present our findings related to students' strategy use. Finally, we will connect the findings back to prior studies with the PSMs and provide opportunities for their use.

Full Name	**Department/Division**	**University/Company/Organization**	**Email Address**
Jonathan Bostic	School of Teaching and Learning	Bowling Green State University	bosticj@bgsu.edu
Gabriel Matney	School of Teaching and Learning	Bowling Green State University	gmatney@bgsu.edu
Timothy Folger	Educational Foundations, Leadership, & Policy	Bowling Green State University	tdfolge@bgsu.edu
Noah Brown	School of Teaching and Learning	Bowling Green State University	bnoah@bgsu.edu

Emily Evans	School of Teaching and Learning	Bowling Green State University	evansea@bgsu.edu
Toni Sondergeld	Educational Leadership & Management and Teaching, Learning, & Curriculum	School of Education	Tas365@drexel.edu
Gregory Stone	N/A	Metriks Amerique	gregory@metriks.com

Deepening the Validity Argument for the Problem-Solving Measures 3-5

Jonathan David Bostic, Gabriel Matney, Timothy Folger, Noah Brown, Emily Evans, Toni Sondergeld, and Gregory Stone

Related literature

Problem solving is a key feature of mathematics (NCTM, 2014, 2000, 1989). Word problems are a central part of mathematics teaching and learning (Bostic et al., 2016; Palm, 2006, 2008). For the purpose of this study, problem solving is defined as, "the process of interpreting a situation mathematically, which usually involves several cycles of expressing, testing, and revising mathematical interpretations" (Lesh & Zawojewski, 2007, p. 782). Problem solving goes beyond the type of thinking needed to solve exercises, which are meant to promote proficiency with a known procedure (Kilpatrick et al., 2001; Mayer & Wittrock, 2006; Polya, 1945/2004). Problem solving occurs when the task is a problem, not an exercise (Schoenfeld, 2011). For the present study, we utilize Verschaffel and colleagues' (2000) problem-solving framework to consider students' sensemaking of word problems. This six-stage framework of problem solving includes: (a) reading the problem, (b) creating a representation of the situation, (c) constructing a mathematical representation of the situation, (d) arriving at a result from employing a procedure on the representation, (e) interpreting the result in light of the situational representation [see (b)], and finally, (f) reporting the solution within the problem's context. This framework dovetails with the way Verschaffel et al. (1999) characterizes word problems as (a) open, (b) developmentally complex, and (c) realistic tasks for an individual. Open problems can be solved using multiple developmentally-appropriate strategies. Problems are developmentally complex for a student when they require productive thinking and are not readily solvable

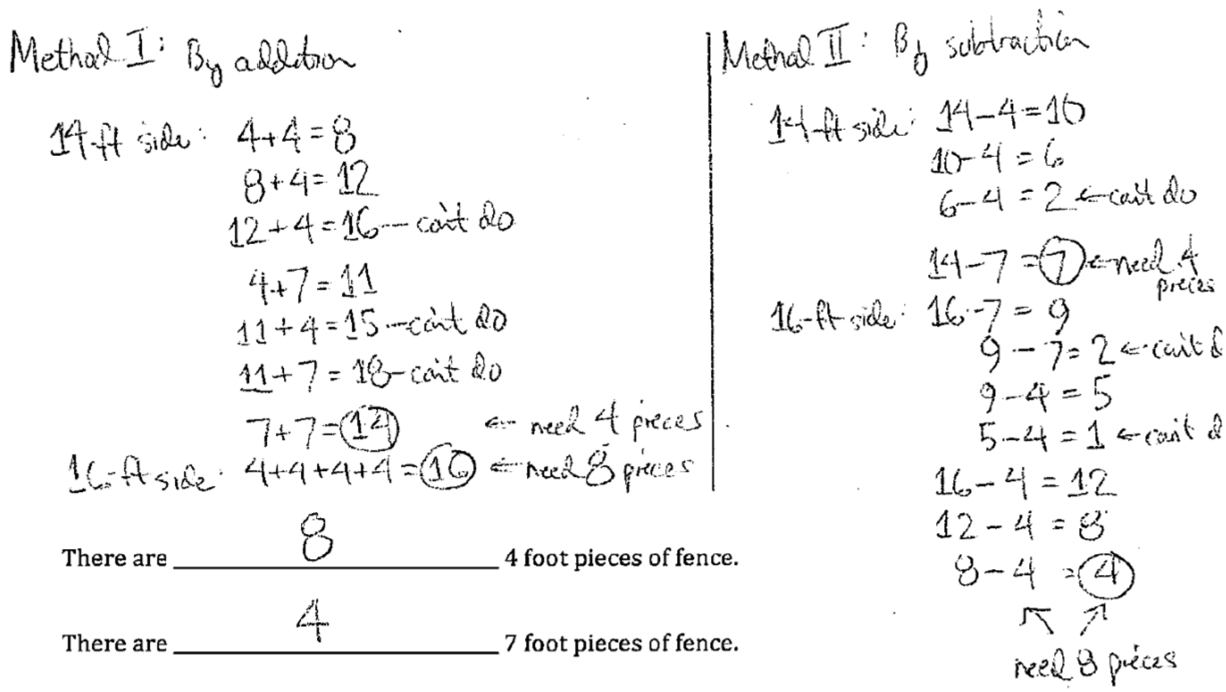
(Schoenfeld, 2011). Realistic tasks may draw upon real-life experiences, experiential knowledge, and/or believable situations (Verschaffel et al., 1999).

A central part of problem solving is strategy use (Verschaffel et al., 2000; Yee & Bostic, 2014). Mathematical strategy use includes both the representations and procedures used to solve problems (Goldin, 2002). Representations include words, symbols, pictures, and even mixed strategy use that combines two representations (Lesh & Doerr, 2003; Matney et al., 2021; Yee & Bostic, 2014). Procedures describe the steps by which a problem solver uses to move from a given task to a result. An example of two different procedures to solve the following task are shown in figure 1 because it can be difficult to imagine unique procedures to solve the same task. The Dog Park Task is one item from the Problem Solving Measures (PSMs).

Figure 1

Different mathematical procedures to solve Dog Park task (PSM3)

Task: The city is creating a dog park. The dog park will be fenced. The space that will be fenced is shown below. Fence pieces measure 4 feet in length and 7 feet in length. Fence pieces may **not** be cut. How many 4 foot pieces of fence are needed? How many 7 foot pieces of fence are needed? Please write your answers on the lines below.



Considering how students solve problems necessitates carefully researching the ways in which they do so on assessments.

Research on Problem Solving Measures (PSMs)

The PSMs are an assessment series that measure problem-solving abilities of students, using the Common Core State Standards for Mathematics (CCSSI, 2011) as the content and practice framework. Development of the PSMs for grades three through eight is grounded in the Standards for Educational and Psychological Testing ([Standards] AERA et al., 2014). The Standards outline five sources of validity evidence: test content, response processes, relations to other variables, internal structure, and consequences from testing/bias. Brief descriptions of the five sources of validity evidence are shared in table 1. The PSM series was designed for use in grades 3-8 (USA) and each grade level is connected to a unique grade-level test (e.g., grade 3 has the PSM3). Figure 1 is a task from the PSM3.

Table 1.
Characterizations of the Five Sources of Validity Evidence

Source of Validity Evidence	**Description**
Test Content	“Test content refers to the themes, wording, and format of the items, tasks, or questions on a test” (AERA et al., 2014, p. 14).
Response Processes	“Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers ” (AERA et al., 2014, p. 15).
Relationship to Other Variables	Relations to other variables may provide evidence, for example, that indicates how “...test scores [may or may not be] influenced by ancillary variables such as [individual or group characteristic]” (AERA et al., 2014, p.12).
Internal Structure	“Analyses of the internal structure of a test can indicate the degree to which the relationships among the items and test components conform to the construct on which the proposed test score interpretations are based” (AERA et al., 2014, p.16).
Consequences of Testing and Bias	“...decisions about test use are appropriately informed by validity evidence about intended test score interpretations for a given use, by evidence evaluating additional claims about consequences of test use that do not follow directly from test score interpretations, and by value judgments about unintended positive and negative consequences of test use” (AERA et al., 2014, p. 21).
Reliability	“The term reliability has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effort on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported” (AERA et al., 2014, p. 33).

Past research has described the development and validation of the PSMs for grades 3-8 (e.g., Bostic & Sondergeld, 2015; Bostic et al., 2017, 2020, 2021; Matney et al., 2021). These validation studies present a strong argument in support of the intended interpretation and use of each PSM. More specifically, the validation of these PSMs highlights the strong connections between the source of validity evidence and its relative contribution to the validity argument. For example, response processes validity evidence were collected through think alouds with test-takers and support the claim that "below-average, average-, and above-average performing students were able to read and solve problems on the PSM6" (Bostic & Sondergeld, 2015, p. 286). However, validation is also an ongoing process (AERA et al., 2014; Cronbach, 1988; Kane, 2016) that must be thoughtfully re-examined continuously. A test is not valid, its interpretations and uses are (Carney et al., accepted; Kane, 2013). A validity argument "encompasses evidence gathered from new studies and evidence available from earlier reported research" (AERA et al., 2014, p. 21). It conveys information to potential test users and administrators about the degree to which the validity evidence supports the interpretations from a test. As such, the current study considers additional validity evidence for the PSM3-5 series based on (1) test content and (2) response processes.

Validity evidence based on test content may be explored through expert panels (AERA et al., 2014). In prior research, mathematicians, mathematics educators, and mathematics teachers reviewed the items and confirmed that the word problems found on the PSMs adhere to a framework for problems related to this study (e.g., complex, open, and realistic; Bostic et al., 2020). While that evidence is indicative that there is good content alignment between the items and desired standards, further evidence can add to the validation argument that the items are functioning as desired from a content perspective.

Validity evidence based on response processes may be collected through cognitive interviews and think-alouds, which elicit students' strategies and/or responses for certain items (AERA et al., 2014). Previous research indicated students view the PSM items as realistic and believable (Bostic et al., 2021). The current research study may add to the validation argument as students' work samples on PSM3-5 items are indicative of problem-solving behavior. More specifically, solution strategies that extend beyond algorithmic procedure provides further evidence that items are open and complex. Thus, we wondered if students' responses were aligned fairly well with the PSMs' construct of interest (i.e., problem solving).

To that end, our research team has examined students' work on these items to identify the degree to which PSM problems actually worked as desired in field testing. The purpose of this study is to examine elementary students' strategy use derived from the context of taking the PSMs as an individual in a testing environment. Use in an actual testing situation might differ from typical response process gathering mechanisms like think alouds or cognitive interviews and we seek to re-engage in the validation process for confirmatory evidence or discriminant evidence. A desired outcome is to add to past content and response processes validity evidence, and consider the degree to which items are functioning as intended. The two research questions for this study are: (a) How do elementary students solve PSM3-5 items? (b) To what degree are multiple strategies used by respondents to solve PSM3-5 items?

Methods

Participants

Our team drew upon a purposeful, representative sample of tests administered to elementary students. This diverse sample spans urban, suburban, and rural school districts within one Midwest state. All children identify as English speaking and do not have disabilities as specified in an Individualized Education Plan.

Data Collection and Analysis

Tests from 21randomly selected teachers across three school districts were selected. This was done intentionally to maintain fidelity through the data analysis. The minimum number of tests for each grade level that were reviewed was 60 tests and a maximum of 120 tests per grade level. The number of items analyzed on each test by grade level were 11, 10, and 10 respectively for grades 3 through 5 for a total of 31 items.

Tests were reviewed by four researchers. Researchers worked in pairs while coding. Pairs observed each students' mathematical work and derived a consensus on the strategies used by participants. Pairs coded students' mathematical work according to previous protocol (e.g., Matney et al., 2021; Yee & Bostic, 2014). The protocol established that students' problem solving strategies be explicitly described and counted according to their use of different representations or different procedures. The results for the present study focus on the frequency of multiple strategies for each problem on a PSM test. The results stemmed from a qualitative data analysis process. Data were coded using inductive analysis (Hatch, 2002). This approach included the following steps, which parallel Hatch (2002) and Creswell (2011) recommendations. (1) Read and become familiar with all tests within a particular grade. (2) Read one PSM problem at a time. (3) Document unique (i.e., representationally or procedurally different) strategies. (4) After documenting all strategies for one PSM problem, examine r similarities and differences across them. (5) Collapse strategies where appropriate into possible groups and characterize each group. (6) Re-examine the strategies for potential misgrouped strategies and/or clarifying characterizations of strategies. (7) Disseminate results from qualitative analysis and communicate any issues during the coding process.

Findings

We examined how elementary students solved PSM3-5 problems (RQ1) and found that students solved the items using different representations and different procedures. We further elucidate these findings via an example from one task on the PSM4, the Fair Task, seen in figure 2.

Figure 2.

PSM4 Fair Task

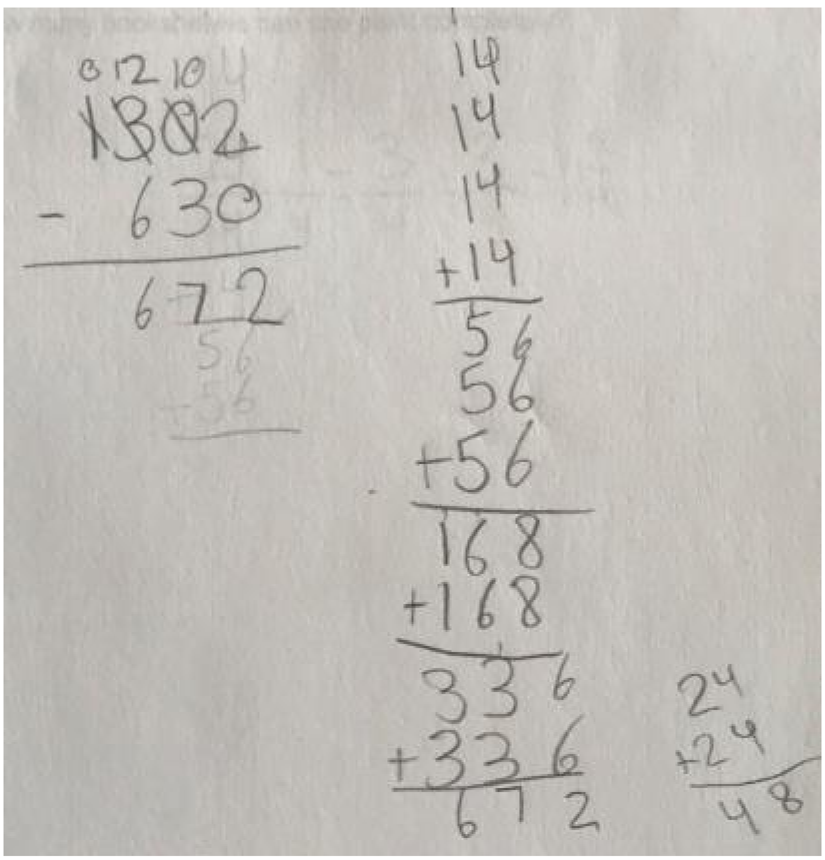
Josephine sold tickets to the fair. She collected a total of $1,302 from the tickets she sold. $630 came from the adult ticket sales. Each adult ticket costs $18. Each child ticket costs $14. How many child tickets did she sell?

The Fair Task asks students to connect three important elements. 1) The difference between the total dollars ($1,302) and the adult ticket dollars ($630) is the amount spent on children's tickets ($672). 2) Each group of $14 dollars spent represents 1 child ticket sold. 3) The number of groups of 14 within the difference ($672) is the number of child tickets.

In the sample of tests we analyzed, students approached solving the Fair Task problem using three unique strategy types. Throughout the three strategies for this task, we noted that all fourth grade students sampled used the standard subtraction algorithm to find the difference on this problem. This finding was not very surprising as the CCSSM holds an expectation for fluency with this algorithm by the end of 4th grade. The standard reads, "4.NBT.B.4 Fluently add and subtract multi-digit whole numbers using the standard algorithm." Students began to diversify their approaches in the less familiar content area involving the number of groups of 14 to determine the number of child tickets. In the first strategy, (see Figure 3) students used repeated addition to find the number of groups of 14 within 672. Students went about this strategy with varying efficiencies. Although it is possible to add 14 to itself 48 times to see the sum is 672, students tended to move toward more efficient additions, keeping track of the number of groups within lesser sums. The student work in Figure 3 demonstrates students' movement toward efficiency as they solved the problem. This particular student began by adding 14 four times and noted that the sum was 56. Then the student noted they could add 56 three times, which gave them 168, or 12 groups of 14. Then the student doubled 168 to get 24 groups of 14, and doubled that again to get 48 groups of 14 and the sum of 672. The student recognized that this was both the desired sum and the corresponding number of groups asked for by the problem.

Figure 3.

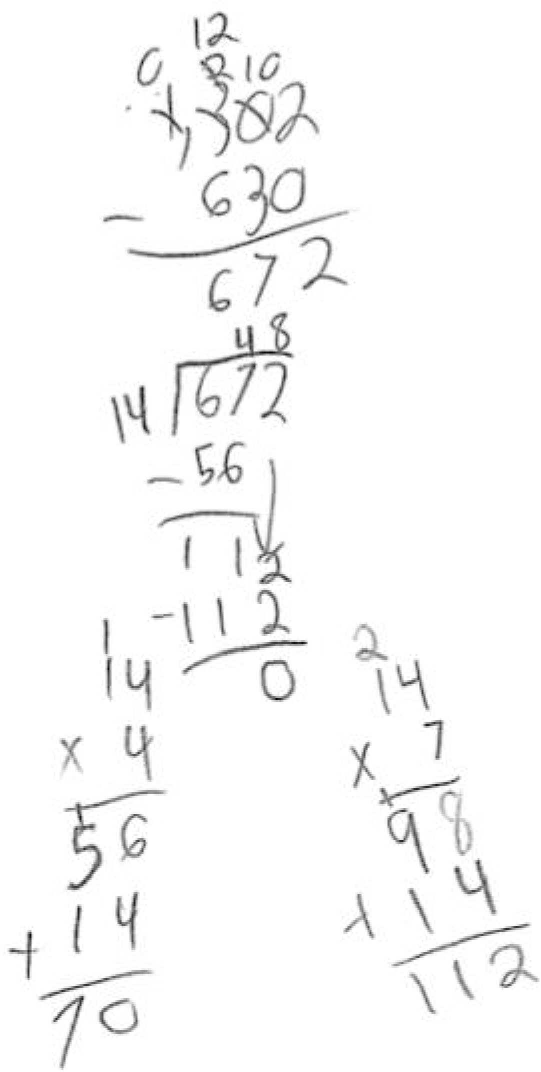
Strategy use with repeated additions to find the number of groups of 14 within the difference of 672.



In the second strategy type (see Figure 4) we observed students using, or attempting to use, the standard division algorithm to find the number of groups of 14 within the difference of 672. The student example shown in Figure 4 shows the common supporting multiplications that students needed to complete their work using the standard division procedure.

Figure 4.

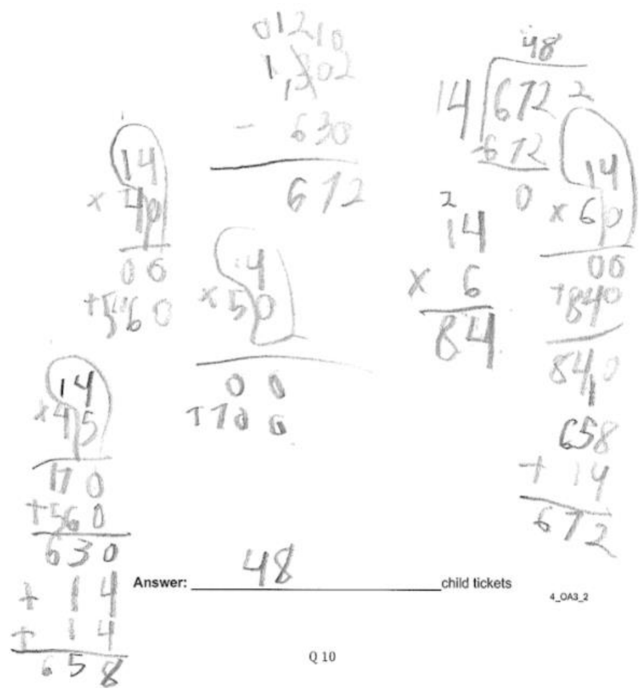
Strategy use with the standard division algorithm to find the number of groups of 14 within the difference of 672.



In the third strategy type for the Fair Task, students used multiplication to guess and check how many groups of 14 within the difference of 672. Similar to repeated addition, student strategies of this type showed differing attempts at efficiency. Some students strategically built up the multiples of 14 using the tens (10, 20, 30, 40) until reaching 48 groups of 14 while others appeared to be randomly guessing at first until they narrowed down the number of groups. The student work shown in Figure 5 shows an advancement of a student's efficiency using tens of groups of 14 (40 and 60) to narrow down how many groups of 14 were needed. The student likely noticed that the needed number of groups was closer to 40 than 60 and then tried 45 groups of 14. Seeing that 45 groups of 14 were close, the student used addition to finish off finding 45+3 groups of 14 was the desired number of groups.

Figure 5.

Strategy use with multiplication to guess and check for finding the number of groups of 14 within the difference of 672.



Similar variations in strategy types being enacted by students were found across all 31 problems for the three PSMs. In consideration of the degree to which students were solving problems in unique ways (RQ2), we collected the frequency with which strategies were found for each problem. Across all problems, the fewest number of strategies generated by students was two and the maximum number of strategies was five. In Table 2, we share the number of items for each grade level in which there were multiple strategies demonstrated by students. Students demonstrated multiple strategies for solving the problems.

Table 2. Frequencies of items demonstrating the range of different strategies from two to five

	Number of Problems Solved with 2 to 5 Different Strategies				**Total Items**
	2 Strategies	3 Strategies	4 Strategies	5 Strategies	
Grade 3	1	5	1	4	**11**
Grade 4	3	3	2	2	**10**
Grade 5	2	5	1	2	**10**

Discussion

Students, generally speaking, provided evidence that their strategy use mirrored experts' recommendations. There is strong evidence that respondents were able to provide multiple strategies to solve PSM items. This provides confirmatory evidence for two validity sources: test content and response processes. There is robust evidence that students' strategy use on the PSMs3-5 aligns with desired outcomes from test content. This evidence implies that there is further support for a claim that the PSM3-5 items address desired content as defined in the test blueprint. Following in the model of Kane (2013) and Carney et al. (accepted), it is reasonable to assert that the PSMs measure mathematical problem solving in ways that are connected to the mathematics described in the CCSSM. This statement about validation provides strength for potential test users to believe that the PSMs measure what they intend.

A second facet drawn from this evidence is that the evidence adds further response process evidence. Problem solvers provided mathematical strategies that reflected past research with think alouds. In addition, mathematicians, mathematics educators, and mathematics teachers provided mathematical strategies that were developmentally appropriate for each PSM3-5 item. Results from this purposeful sample of a large-scale administration of the PSMs that those respondents seem to be using the strategies predicted by expert panel members and think aloud respondents. It can be logically inferred that the PSM3-5 items are eliciting a desired response from test takers. More specifically, they are using a diverse array of developmentally appropriate strategies to solve the word problems.

An implication from this research is that there is deeper evidence grounding the validation argument for the PSMs. That is, the PSMs are intended to measure the degree to which students are able to solve mathematics problems that are grounded in the mathematics content that are used in many school districts. As a result, PSM users may feel confident because not only was there sufficient evidence that respondents could solve PSM problems in multiple ways supported by content expert and think aloud analysis, but additionally the present study adds to this validity evidence demonstrating that students continue to show multiple ways of solving PSM problems in the normal testing environment. The validity evidence from the present study fuels the claim, driven by test content and response process as a frame, that the PSMs are written in a way that promote respondents to use multiple strategies that are developmentally appropriate.

Future Research and Conclusion

Further research is warranted to examine how strategies differ across items and potential areas where there exists misconceptions. Another area is to examine the strategy use of students who have an identified disability or are identified as English Language Learners. While the PSMs are typically administered to K-12 students, it may be worthwhile to explore how teachers (preservice or inservice) solve PSM problems. Taken collectively, these areas of research would support further evidence about what respondents know and are able to do on an instrument designed for broad use. PSM users and administrators may feel confident that the PSMs measure problem solving and mathematics performance in a robust, valid manner.

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., 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., Pape, S., & Jacobbe, T. (2016). Encouraging sixth-grade students' problem-solving performance by teaching through problem solving. Investigations in Mathematics Learning, 8(3), 30-58.

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., 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 standards for mathematics. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Creswell, J. (2011). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th ed.). Boston, MA: Pearson.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), Test Validity (pp. 3-17). Erlbaum

Goldin, G. (2002). Representation in mathematical learning and problem solving. In L. English & M. Bartolini Bussi (Eds.), Handbook of international research in mathematics education (2nd ed., pp. 176-201), New York: Routledge.

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73.

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), Handbook of test development. (2nd ed., pp. 64-80). Routledge/Taylor & Francis Group.

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

Lesh, R., & Doerr, H. (2003). Foundations of a models and modeling perspective on mathematics teaching, learning, and problem solving. In R. Lesh and H. Doerr (Eds.), Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teachin g (pp. 3-33). Mahwah, NJ: Erlbaum.

Lesh, R., & Zawojewski, J. (2007). Problem solving and modeling. In F. Lester, Jr. (Ed.), Second handbook of research on mathematics teaching and learning (pp. 763-804). Charlotte, NC: Information Age Publishing.

Matney, G., Bostic, J., Fox, M., Sondergeld, T., & Stone, G. (2021, March). Fourth-grade students sensemaking of word problems. In H. Marchionda & S. Bateiha (Eds.), Proceedings of the 48th Annual Meeting of the Research Council on Mathematics Learning (pp. 59-66). Denton, TX.

Mayer, R., & Wittrock, M. (2006). Problem solving. In P. Alexander & P. Winne (Eds.), Handbook of educational psychology (pp. 287-303). Mahwah, NJ: Erlbaum.

National Council of Teachers of Mathematics. (2014). Principles to action. Reston, VA: Author.

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

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

Palm, T. (2008). Impact of authenticity on sense making in word problem solving. Educational Studies in Mathematics, 67, 37-58.

Palm, T. (2006). Word problems as simulations of real-world situation: A proposed framework. For the Learning of Mathematics, 26, 42-47.

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

Reed, S. K. (1999). Word problems: Research and curriculum reform. Mahwah, NJ: Erlbaum.

Schoenfeld, A. (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.

Verschaffel, L., Greer, B., & DeCorte, E. (2000). Making sense of word problems. Lisse, Netherlands: Swets & Zeitlinger.

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