PSM5: Measuring Elementary Students' Mathematical Problem Solving

Problem solving is central to mathematical work and is a core component of mathematical standards found in many instructional standards. As such, effective measures must be available to scholars and school personnel. Assessment is a key part of the teaching and learning process. Additionally, scholars need rigorous assessments with robust validity evidence to use results from those assessments as part of generalizable research. The focus of this paper is development of one problem solving measure for grade 5 (ages 10-12) students. It is a measure within the Problem-Solving Measure (PSM) series. Results from this study offer a quantitative instrument that can be used broadly.

Assessment and instruction should be connected (Black & Wiliam, 1998). Problem solving is found in instructional standards for elementary and secondary students as described in the Standards for Mathematics Content (SMCs). For example, fifth-grade students are expected to "relate volume to operations of multiplication and addition and solve real world and mathematical problems involving volume" (5.MD.5, p. 37). Similarly, "make sense of problems and persevere in solving them" is the first Standard for Mathematical Practice (SMP; CCSSI, 2010, p. 6). Problem solving is also found in mathematics standards from many countries around the world thus it is a global concern (Mullis et al., 2016). Mathematical problem solving continues to be a lynchpin for further success in Science, Technology, Engineering, and Mathematics (Committee on STEM Education, 2018). It is a high priority within instructional standards around the world as seen in both process/practice and content standards (Mullis et al., 2016). Mathematics instruction in K-12 schools has had an increased focus on processes and practices directly related to problem solving since 1960 (Li & Schoenfeld, 2019). Problem solving is noted in mathematical standards that starts with instruction for elementary students and continues through high school (see CCSSI, 2010), which means that problem-solving assessments for elementary students should parallel mathematics instruction. If mathematical problem solving is expected to be part of classroom instruction that addresses standards for elementary students, then how might it be assessed in ways that reflect modern expectations for validity and reliability? The purpose of this study is to describe one test for fifth-grade students that is part of a series of the Problem-solving Measures (PSM) designed for grades 3-8. An overarching research question guiding this study is: What validity evidence supports the use of the PSM5?

Related Literature

For 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 can only happen when students are engaged in problems. Problems, for the present study, are defined using two frameworks. First, the solution strategy for a task is uncertain at first glance, there exists more than one way to complete it, and the solution (or number of solutions) is unknown (Schoenfeld, 2011). A second framework characterizes problems as open, complex, and realistic tasks (Verschaffel et al., 1999). Open tasks can be solved in more than one way. Complex tasks encourage problem solvers to think critically about the mathematics they need to work on. Realistic tasks draw upon contexts where individuals use experiential knowledge in ways that connect in- and out-of-school knowledge. Problems, including word problems, are different from

exercises. Exercises are tasks meant to foster students' efficiency with a known procedure (Kilpatrick et al., 2001). To that end, this study intends to examine the qualities of a problem-solving assessment.

Assessments are expected to adhere to *Standards for Educational and Psychological Testing* (AERA et al., 2014). These *Standards* recommend gathering validity evidence for up to five sources of validity: test content, response processes, relationships to other variables, internal consistency, and bias/consequences from testing (see Table 1 for definitions). Validity evidence helps to convey to others how certain one can be of results and interpretations from the assessment (Kane, 2016). It is not necessary to gather evidence for all five sources of validity; but more robust validity arguments are more likely to have evidence from multiple sources and/or numerous pieces of evidence (AERA et al., 2014; Author, 2019). Intellectual merit of results from using high quality measures leads to implications that have stronger implications for research and practice (Author, 2019). Results stemming from assessments lacking validity evidence can lead to spurious findings. Additionally, instrument developers should are expected to provide an instrument use summary that communicates what an instrument can do, how it should be used, and evidence for those statements (Author, accepted). However, this has been done rarely in practice (Author, accepted, 2020, 2019), which has led to many challenging issues. This validation study begins to fill a gap in mathematics education assessment literature with information how to use the PSM5 appropriately, which comes from a well-documented validation study.

Method

Context and Participants

This study draws upon quantitative and qualitative data to communicate a validity argument for the PSM5, resulting in a measure that has potential for uses within the USA. The validity study is framed by the five sources (see AERA et al., 2014), which has been used with prior PSMs (see Author, 2015, 2017). For test content, we assembled an expert panel consisting of grade 5 teachers, terminally degreed mathematics educators whose background is working with elementary students and teachers, and university-level mathematicians. Response processes evidence was gathered from 56 purposefully selected fifth-grade students who participated in think alouds. Students were selected along different variables: ethnicity, gender, and past achievement performance. Relations to other variable evidence and internal consistency data used 373 students' PSM5 responses. Students were native English speakers and came from rural, suburban, and urban schools across a Midwest state. Bias/consequences from testing evidence came from eight purposefully selected student interview volunteers following test administration. Additionally, six fifth-grade teachers offered feedback about potential areas of bias as well as their perceptions of students' affect following test administration.

Instrumentation

The PSM5 has 12 word problems and each is presented as a constructed response task. Students are asked to show their work and clearly write their answer on a provided line. The target population are English-speaking, grade-level appropriate students. A Flesch-Kincaid (Kincaid et al., 1975) readability analysis indicated that the PSM5 items meet grade-level readability expectations (PSM5 = 5.1). PSM5 administration is typically performed during instruction and may take up to 120 minutes; however, most finish within 75-90 minutes. There is no difference in student outcomes whether the PSM5 is completed in one sitting multiple sittings

(e.g., six 20-minute administrations). Calculators are not allowed during PSM5 administration. An example of a PSM5 item is in figure 1.

Drake's parents served cake at a party. $\frac{2}{3}$ of the cake was eaten. Terence and Sean came over the next day because they did not attend the party. Sean and Terence ate the remaining cake equally. What fraction of the original cake did Terence eat?

Figure 1. *Sample PSM5 Item.*

Data Collection and Analysis

Qualitative data were collected for test content, response processes, and bias/consequences from testing. Expert panel members reviewed items for connections to grade-level standards, and the degrees to which items (a) were complex, (b) had a verifiable solution set, and (c) could be solved in multiple ways that are developmentally appropriate for respondents. Students participating in think-alouds for response processes evidence were asked to share their thinking aloud as they worked on the items. A researcher jotted notes while the students worked and collected students' work after all items had been administered. Bias/consequences from testing data were gathered during interviews with expert panel members as well as students who completed the PSM5. These expert panel, response processes, and bias/consequences from testing data were analyzed using inductive analysis (Creswell, 2012; Hatch, 2002). The purpose of inductive analysis was to identify salient themes from data.

Quantitative data were collected for relations to other variables, internal consistency, and bias/consequences from testing. Each PSM5 item was scored dichotomously, which conveys the same information as partial credit scoring (Author, in press). Respondents' scores may be calculated as percent correct. For relations to other variables evidence, data for ethnicity (white/nonwhite), gender, and prior achievement were gathered. Quantitative data (relations to other variables, internal structure, and consequences from testing) were analyzed using Rasch (1960/1980) modeling and traditional statistics. Rasch is commonly used to assess differential item functioning (DIF) based on gender and race/ethnicity (Bond & Fox, 2007), which can be used as an indicator of consequences from testing. Person measures are stable within 0.5 logit with 95% confidence with sample sizes between 64 and 144 respondents (Wright & Stone, 1979). Separation and reliability scores of 2.00 and .80 are considered good while 3.00 and .90 are excellent (Duncan et al., 2003). Finally, negative point biserial values indicate issues such as higher performing students more likely to respond incorrectly and lower performing students to respond correctly to an item. It is important to note that problem solving is more difficult than completing procedures; thus, average student ability (i.e., person means) are expected to be low (Author, 2015, 2017; Verschaffel et al., 1999). Rasch reliability and traditional reliability statistics (i.e., Cronbach's alpha) are similar in that they both describe the statistical reproducibility of a set of values.

Results

Validity evidence for the PSM5 and the associated claims is grouped by the five validity sources, which is shown in Table 2. We summarize the results here due to word limits and will provide details in our presentation. Results from analyzing test content data indicated agreement across the expert panel that items were aligned with grade-level content; were complex in nature; had a verifiable solution set; could be solved in at least two developmentally-appropriate ways; and involved contexts that were realistic to students. Results from analyzing response processes

data suggested students responded in anticipated ways. Not every student arrived at the correct solution for each item but they implemented anticipated strategies noted by the expert panel. Results from analysis of relations to other variables conveyed that there were no statistically significant relationships between PSM5 outcomes and gender or ethnicity status. There was a statistically significant relationship between prior achievement and PSM5 outcomes (e.g., high prior achievement and greater PSM5 score). Internal consistency results showed that item separation (7.02) and reliability (.98) were excellent. Respondents' mean score was -1.1 logits. No items had negative point-biserial values. Finally, consequences from testing/bias results were clear: Students reported neutral or positive affect following test administration and did not express feeling any bias within the items. In fact, many were excited to solve realistic problems. Expert panel members did not perceive any implicit bias in the items. Classroom teachers indicated students behaved similarly after the PSM5 administration as they do after a unit test.

Discussion and Importance

The PSM5 validity evidence addresses the five sources and there is justification for its use. These pieces of evidence and their associated claims provide assurances that results and interpretations from the PSM5 are appropriately linked with the measure. Problem solving is an important topic found across content and practice standards; thus, students' problem-solving outcomes must be assessed in ways that lead to valid interpretations and findings that researchers and school personnel can trust. The PSM5 provides scholars and school personnel with measures that can be used at scale. PSMs are designed to complement other data about students' mathematics outcomes and be interpreted as a single touchpoint of students' outcomes. They are not intended as a placement tool or to rank students and their teachers. 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).

Limitations and Future Directions

First, this research includes only native English-speaking students and further research is warranted to explore outcomes with students who do not speak English as their primary language. Data from Hispanic and Latinx students was limited; hence, we will gather more data from students representing this growing population in a future study. Second, we are currently working with intervention specialists and students with disabilities to explore how the PSM5 may be used as an assessment tool to support each and every child's mathematical learning. Third, we are actively researching vertical equating across the PSM series so that each test is situated on the same logit ruler. This allows for the PSMs to accurately reflect students' growth as mathematical problem solvers from one grade level to the next.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Author. (accepted).

Author. (in press).

Author. (2020).

Author. (2018).

Author. (2017).

Author. (2019).

Author. (2015).
Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of General Internal Medicine, 20*(12), 1159-1164.
Black, P & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in education: Principles, policy & practice,* 5(1), 7-74.
Bond, T., & Fox, C. (2015). *Fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
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.). Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.
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, A. (2002). *Doing qualitative research in education settings.* Albany, NY: State University of New York Press.
Kane, M. T. (2016). Validation Strategies: Delineating and Validating Proposed Interpretations and Uses of Test Scores. In S. Lane, M. Raymond & T. M. Haladyna (Eds.), *Handbook of Test Development* (Vol. 2nd; pp. 64-80). New York, NY: Routledge.
Kilpatrick, J., Swafford, J., and Findell, B. (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.
Lesh, R., & Zawojewski, J. (2007). Problem solving and modeling. In F.K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 763-803). Charlotte, NC: Information Age.
Mullis, I. V. S., Martin, M. O., Goh, S., & Cotter, K. (Eds.) (2016). TIMSS 2015 Encyclopedia: *Education Policy and Curriculum in Mathematics and Science.* Retrieved from http://timssandpirls.bc.edu/ timss2015/encyclopedia/
Padilla, J. L., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136-144.
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. https://doi.org/10.7334/psicothema2013.256

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.
Wright, B.D., & Stone, M. (1979). *Best test design.* Chicago, IL: MESA Press.

Table 1
Validity Evidence Types Operationally Defined with Aligned Data Sources

Validity Evidence	Operational Definition	Typical Supporting Evidence
Test Content	Instrument item alignment (test content) with the construct to be measured (theoretical trait).	Subject matter experts (SMEs) evaluating item-to-construct alignment and can be logical or empirical (qualitative) (Sireci & Faulkner-Bond, 2014).
Response Process	Participant responses or performance alignment with the assessment construct.	Cognitive interviews, think alouds, or focus group interviews, using a sample of typical respondents to verify that they interpret items and respond in ways developers imagined they would (qualitative) (Padilla & Benitez, 2014).
Internal Structure	Extent to which items and components of instrument reflect the construct.	Psychometric related to: 1) instrument dimensionality, 2) measurement invariance, and 3) instrument reliability (quantitative) (Rios & Wells, 2014).
Relationship to Other Variables	Instrument outcome associations with other variables hypothesized to be related (either positively or negatively).	Statistical testing between instrument outcomes and potentially associated variables (quantitative) (Beckman et al., 2005).
Consequential	Negative impact from completing assessment or item/ instrument bias.	Participant perceptions of instrument impact on them (qualitative) (Authors, 2015).

Table 2.
Connecting Validity Evidence and Claims for PSM5

Data source	Data Analysis Approach	Validity source	Validity Claim
Expert panel	Qualitative: Inductive analysis	Test content	PSM5 items address the mathematics content and practices described in the standards (CCSSI, 2010)
Expert panel; 1-1 think alouds and whole-class think alouds	Qualitative: Inductive analysis	Consequences from testing/bias	PSM5 items have appropriate face validity and do not promote bias in favor of one group of students over another.
1-1 think alouds and whole-class think alouds	Qualitative: Inductive analysis	Response process	Students respond to PSM5 items in anticipated ways. Students perceive PSM5 items as being realistic and relate to the word problem contexts.
Administration with large sample	Quantitative: Rasch analyses	Internal structure (and reliability)	PSM5 consists of an item set that conforms to a single construct called mathematical problem solving.
Administration with large sample	Quantitative: Rasch analysis	Relations to other variables	PSM5 scores are related to respondents' mathematical ability (e.g., students with high math achievement are more likely to have higher PSM5 scores than those of average math achievement). There is no relationship between PSM5 scores and gender or ethnicity.