Developing and Evaluating Assessments of Problem Solving (DEAP) : Year 1

Jonathan Bostic (PI), Gabriel Matney (co-PI), Toni Sondergeld (co-PI), and Gregory Stone

Bowling Green State University, Drexel University, and Metriks Amérique

bosticj@bgsu.edu; gmatney@bgsu.edu; tas365@drexel.edu; gregorystone@metriks.com

Research Focus for Year 1

To what degree does validity evidence support use of the Problem-Solving Measure (PSM) grades 3, 4, and 5 to measure students' problem-solving abilities related to the mathematics content and practices described in the Common Core State Standards? We conducted steps 1, 2, and 3 of the validation process during year 1, which is shown in Figure 1.

Purpose

There are three aims of DEAP. (a) Create three new PSMs (grades 3, 4, and 5) and gather validity evidence for their use. (b) Link new PSMs with the already functioning middle-school PSMs (grades 6, 7, and 8). (c) Construct a reporting system and investigate how the reporting system formatively informs teachers' instructional decisions.

Previous Work

Previously, we created the PSMs for middle school students (see Bostic & Sondergeld, 2015; 2018; Bostic, Sondergeld, Folger, & Kruse, 2017). Tests followed the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) as a frame for gathering validity evidence. The five sources of validity evidence (see Table 1) are (1) test content, (2) response processes, (3) relations to other variables, (4) internal structure, and (5) consequences from testing. Grades 7 and 8 tests were vertically equated (linked) with the test preceding it (grade 6 and 7, respectively).

Sample Item from PSM4

"A group of 96 tourists waited in a parking lot for a boat to take them to an island. The boat can carry 7 people on each trip. After a few hours, everyone in the group of 96 tourists visited the island. What is the fewest number of trips to the island made by the boat?"

Figure 1. Validation process

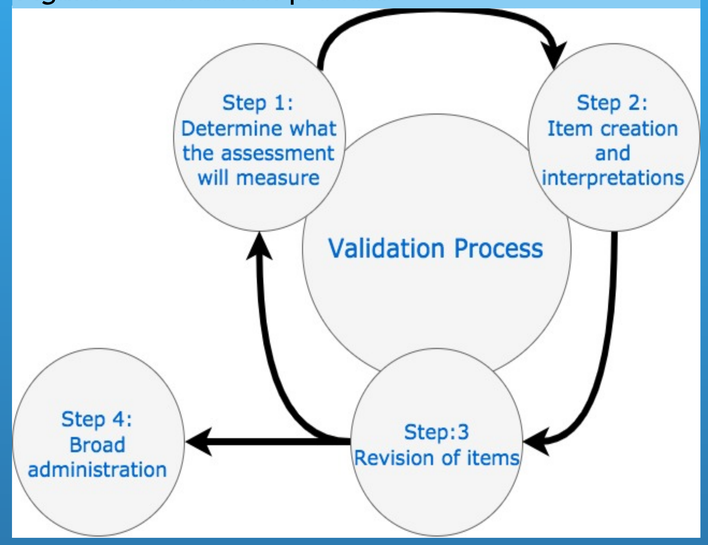

Table 2. Source of validity and evidence collection

Validity Source	Evidence Gathered	Who/What involved
Test Content	Expert Panel	Mathematics teachers, early childhood mathematics educators, and mathematicians
Response Processes	Think-aloud data	Students nested in multiple classrooms within each grade level during April 2018
Relations to Other Variables	Pilot test data	Current academic ability and ethnicity
Internal Structure	Pilot test data	Cronbach's alpha and Rasch reliabilities
Consequences from Testing	Think-aloud data	Students nested in multiple classrooms within each grade level during April 2018

Results and Future Implications

1. Validity evidence suggests that students' outcomes on the PSM3, PSM4, and PSM5 are indicating respectable validity evidence (see Table 2). We intend to conduct further think alouds and conduct larger test administrations in 2018-2019.
2. Teachers have shared positive impressions of the PSM3, PSM4, and PSM5 during think-aloud administration. Many expressed that watching the think-aloud indicated what content to focus on for future instruction. Thus, like the PSM6, PSM7, and PSM8, the PSMs for elementary school have potential to serve as formative assessment tools to guide teachers' instruction.
3. PSMs (3-8) have potential to be used by school districts and education researchers to measure students' mathematics outcomes. Those interested in the PSMs should contact the PI (bosticj@bgsu.edu).

Table 1. Description of five sources of validity

Sources of Validity	Brief Description
Test Content	This source ensures that the assessment is actually a measure of the construct (Lavery et. al., 2017; Cureton, 1951; Kane, 2012). It also takes a deeper look at the question and compares it to the domains that are presented in state standards. It ensures that the questions are of high cognitive level and that the questions assess the most important aspects of the domain (Sireci & Faulkner-Bond, 2014).
Response Process	This source analyzes how participants might react to the item. It ensures that the interaction between the item and the participant is as desired. This evidence expresses how students engage with the items, but it can also be used to answer questions about why different groups perform better on the test than others (AERA et al., 2014).
Internal Structure	This source analyzes items to determine that they accurately correspond to the intended construct of the test (AERA et al., 2014). It also investigates what information the item can provide, determine if there is any bias, and also to ensure the test is written in a way that is reliable.
Relations to Other Variables	This source analyzes the relationships between the measure of interest and other variables. (Lavery et. al., 2017) Evidence can be convergent, meaning there is a relationship, or discriminant, meaning there is not a relationship between the measure of interest and other variables (AERA et al., 2014).
Consequences of Testing	This source analyzes the possible interpretations that may come from the assessment. There are certain questions that may be asked the can make the participant upset, uncomfortable, or even happy and confident. The consequences are typically unintended and can be either positive or negative. This should be explored during test development and again following test use.

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. (2018). In D. Thompson, M. Burton, A. Cusi, & D. Wright (Eds.), *Validating and vertically equating problem-solving measures. Classroom Assessment in Mathematics: Perspectives from Around the Globe*, pp. 139-155. Cham, Switzerland: Springer.

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.

Cureton, E. E. (1951). In E. F. Lingquist (Ed.), *Educational measurement* (pp.621-694). Washington, DC: American Council on Education.

Kane, M. (2012). All validity is construct validity. or is it? *Measurement: Interdisciplinary Research And Perspectives, 10*(1-2), 66-70.

Lavery, M. R., Holloway-Libell, J., Amrein-Beardsley, A., Pivovarova, M., & Hahs-Vaughn, D. (2016). *Evaluating the validity evidence surrounding the use of student standardized test scores to evaluate teachers: A centennial, systematic mega-review*. Paper presented at the American Educational Research Association Annual Meeting, Washington, DC.

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

Acknowledgment: This material is based in part on work supported by the National Science Foundation (DRK-12 Grant #1720646 and #1720661). Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation. We wish to thank Davis Gerber who served as the DEAP research assistant during Year 1.