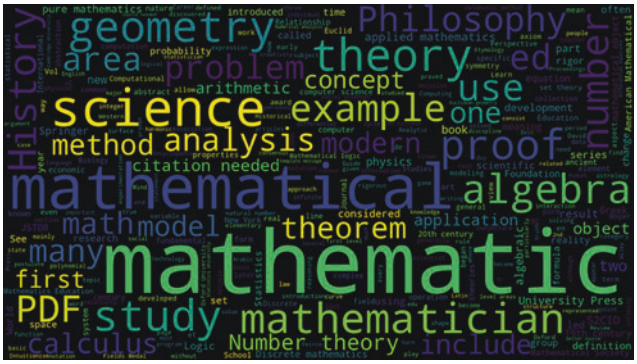
Proceedings for the 52nd Annual Meeting of the Research Council on Mathematics Learning

Celebrating 52 years of Research on Mathematics Learning



Innovating and Integrating: Advancing Mathematics Learning Across Disciplines

March 6 – 8, 2025
College Station, Texas

INVESTIGATING DIFFERENCES IN ASSESSMENT DELIVERY FORMATS: AN ILLUSTRATION STUDY

Toni A. May
Binghamton University
tmay3@binghamton.edu

Gregory E. Stone
Metriks Amerique
gregory@metriks.com

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

Timothy Folger
Binghamton University
tfolger@binghamton.edu

Connor J. Sondergeld
Metriks Amerique
c.sondergeld@metriks.com

This study explored how mathematics problem-solving constructed-response tests compared in terms of item psychometrics when administered to eighth grade students in two different static formats: paper-pencil and computer-based. Quantitative results indicated similarity across all psychometric indices for the overall tests and at the item-level.

Our research team has developed and validated a series of paper-pencil, vertically equated, mathematical problem-solving measures for grades 3-8 called Problem Solving Measures (PSMs 3-8) and shared findings from prior validation studies (e.g., Bostic & Sondergeld, 2015; Bostic et al., 2017). Each PSM was designed to align with the Common Core State Standards for Mathematics (CCSSI, 2011). To expand past scholarships, we began a multi-year process of developing and validating new items for a computer adaptive testing environment. Bostic and colleagues (2024) outline a validation study for the computer adaptive (CAT) mathematical problem-solving measures, which we call DEAP-CAT. During validation and development, we realized that general research on the comparability of results from paper-pencil and computer-based test formats focused primarily on multiple-choice questions and results varied depending on testing contexts (Hamhuis et al., 2020). Further, there was a dearth of research comparing assessment psychometric properties between formats specifically related to mathematical problem solving. Thus, the purpose of this study was to psychometrically compare mathematical problem-solving constructed response item assessments using the same items administered to 8th grade students in paper-pencil and static computer-based formats.

Relevant Literature

Comparing Assessment Delivery Formats

Research on the effect that assessment delivery format (i.e., paper-pencil vs. computer-based) has on testing results has yielded contrasting findings based on specific contexts, including content area, grade level, and item type (Hamhuis et al., 2020; McClelland & Cuevas, 2020;

Puhan et al., 2007). A testing *mode effect* refers to "the likelihood of differential student performance due to differences in how items are presented in [paper-pencil tests] versus [computer-based tests]" (Hamhuis et al., 2020, pp. 2341-2342). At times, research has shown that students perform better on paper-pencil tests compared to computer-based (e.g., McClelland & Cuevas, 2020; VanDerHeyden et al., 2023). However, other research found no significant difference in student performance based on testing medium (e.g., Hamhuis et al., 2020; Threlfall et al., 2007). Further, research has suggested that the existence of a testing mode effect may depend on individual students' backgrounds and characteristics (Hamhuis et al., 2020).

When specifically investigating mathematical constructed-response assessments, one study showed sixth-grade students performed better when the test was delivered in paper-pencil format rather than computer-based (McClelland & Cuevas, 2020). By taking a deeper look at *how* students engaged with mathematical word problems via paper-pencil and computer-based test formats, some research has shown students use different processes (Lemmo & Mariotti, 2017). These results imply that even if student performance in the aggregate is similar across testing mediums, it may not be appropriate to make comparisons of student performance across paper-pencil and computer-based tests (VanDerHeyden et al., 2023). As such, VanDerHeyden and colleagues (2023) concluded that "reliability for the [early-childhood arithmetic test] is only established within each assessment format…but a score obtained in computer-based conditions could not be generalized to scores obtained under paper/pencil conditions and vice versa" (p. 98). While this seems to be a budding line of inquiry, in general, there is a scarcity of research comparing psychometric properties of test items (e.g., difficulty measures, standard error, reliability, fit indices) when the same items are administered in both paper-pencil and computer-based formats, particularly for mathematics problem-solving constructed response items.

Mathematical Problem Solving

Similar to our prior testing scholarship, our research team drew upon two related frameworks for mathematical problems. One frame is that a mathematical problem is a task presented to an individual such that (a) it is unclear whether a solution or how many exist and (b) the pathway to a solution is uncertain (Schoenfeld, 2011). This framing is useful but is not comprehensive for word problem research. Hence, we draw from Verschaffel and colleagues (1999) framing for mathematical word problems as tasks presented to an individual that are open, complex, and realistic. Open tasks may be solved using multiple developmentally-appropriate strategies.

Complex tasks are not readily solvable by an individual and require productive thinking. Open and complex are connected with Schoenfeld's framing of problems. Realistic word problems draw upon real-life experiences, experiential knowledge, and/or believable events. This notion of realism adds a necessary element to effectively frame word problems for our assessment. As a contrast, mathematical exercises are mutually exclusive from problems and are intended to support building an individual's efficiency with a known procedure (Kilpatrick et al., 2001).

Given these two synergistic frameworks for the CAT items and ensuing test, we chose Lesh and Zawojeski's (2007) problem-solving framework for PSM mathematical problem-solving computer adaptive test, which reflects our past test development. Problem solving 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). Problem solving is something that takes time and concentrates goal-oriented efforts on a problem (Polya 1945/2004; Schoenfeld, 2011), which differs from completing exercises.

Method

This research is part of a large federally funded, multi-year initiative to develop and validate items for use in grades 6-8 computer adaptive problem-solving tests. We drew on a design science approach (Middleton et al., 2008) due to its effectiveness in creating assessments through a cyclical process of designing, testing, evaluating, and reflecting. The current study fits into the design science approach by testing comparability findings when PSMs were administered in paper-pencil and static computer-based formats and then reflecting on results and usability.

Participants & Instrumentation

Multiple school districts from three states in the USA representing different geographical regions (i.e., Midwest, Mountain West, and Pacific), varying contexts (i.e., urban, suburban, and rural), and the uniqueness of students' gender and ethnicity were purposefully selected for the larger project. Data from 8th grade mathematics students from those states were specifically used in this study. The samples were not identical across test administrations (because testing was anonymous), but both tests were delivered in the same schools with the same classroom teachers to maintain proximal consistency. Samples for these comparisons were 656 for paper-pencil and 490 for computer-delivered. Final sample sizes ensured we met a minimal item exposure of 30 students per item to properly calibrate performance and ensure statistical performance viability.

In addition, students identified as having a special needs or accommodation (e.g., extra time or tests being read to them) were excluded, to control for this potentially confounding variable. Only fundamental demographic data including gender- and racial/ethnic-identities were gathered and presented in Table 1.

Table 1

Final Sample Student Demographic Characteristics

Student Demographics Values	**Testing Format** **Paper-Pencil (*n*=656)**	**Computer-Based (*n*=490)**
Gender		
Female	232 (35%)	145 (30%)
Male	398 (61%)	335 (67%)
Other	4 (1%)	2 (1%)
Not Reported	22 (3%)	8 (2%)
Racial/Ethnic-Identity		
American Indian/Alaskan Native/First Nations	7 (1%)	6 (1%)
Asian	9 (1%)	10 (2%)
Black or African-American	12 (2%)	7 (1%)
Hispanic/Latino-a or Spanish Origin	48 (7%)	34 (7%)
Middle Eastern or North African	1 (1%)	0 (0%)
Native Hawaiian or Pacific Islander	7 (1%)	5 (2%)
White	551 (84%)	410 (84%)
Other	12 (2%)	7 (1%)
Not Reported	9 (1%)	11 (2%)

Our team sought to develop 240 CAT items for each grade level (i.e., grades six, seven, and eight). After numerous reviews during the item development phase of the project, a total of 182 items associated with 8th grade mathematics content standards met expectations for testing with students. A sample 8th grade item addressing Number Sense CCSSM standards is provided to contextualize the word problems created for the CAT PSMs: "A chess board is made of eight rows with eight squares in each row. Each square has an area of 3 inches^2. What is the exact length for one edge of the chess board?" Similar to past paper-and-pencil PSMs, the CAT PSMs are scored dichotomously.

Data Collection and Analysis

Tests for each delivery format were created using the same bank of 203 previously calibrated problem-solving items. All items in the bank were deemed functional during previous statistical evaluations and linked to one of the five content domains within 8th grade. To ensure

comparability of item calibrations both *across* delivery models (i.e., paper-pencil, computer-based) and *within* each delivery model, a common item equating process using linking items was employed (Kedlermen, 1988). Linking items represent previously calibrated items that are consistent across all versions of the test to ensure that the calibration of items and person abilities are equivalent. Further, using Rasch (1960/1980) modeling places all items on the same linear scale. Linking ensures direct comparability of results and performance statistics.

Paper-pencil tests were designed to be completed in a single 30-minute period, for ease of administration in the classroom. Forty-four versions of such tests were constructed, each consisting of one common item (used for equating) and three to four additional unique items. Each test covered at least four of the five standard domains within 8th grade. Each student, within a class period, took an identical paper-pencil test. Students were able to make use of classroom-provided calculators and scratch paper as needed during the administration.

Computer-administered static-tests were designed and delivered through the FastTest System© (Assessment Systems Corporation, 2023) online under the same conditions used for paper-pencil administration. The full bank of items was entered into the FastTest System© and a set of 44 identical tests were generated. Each test mirrored the features of the original versions. To ensure integrity between paper-pencil and computerized test versions, items that included fractions, square roots, mathematical equations, diagrams, graphs, charts, and pictures were entered into the FastTest system as JPGs. This allowed students to see the same structurally formatted item regardless of test administration format. While students could not write on their computer screen apart from typing their response in a designated response box, they were allowed to use scratch paper for their work, if desired. A classroom-supplied calculator or an electronic calculator embedded in the examination were available for students. The embedded calculator was small enough to fit in the upper corner of the screen without blocking, covering, or hiding any element of the item or its accompanying graphics. Item exposure requirements were applied and the final comparison included 11 tests common to both delivery methods.

Rasch (1960/1980) measurement for dichotomous responses was employed to conduct psychometric analyses for both research questions in this study using Winsteps software (Linacre, 2024). Rasch measurement has long shown its effectiveness in social science instrument development and validation (see Bond & Fox, 2007). Multiple psychometric indices were investigated. Rasch reliability is a measure of internal consistency (acceptable ≥ 0.70, good

≥ 0.80, excellent ≥ 0.90; Duncan et al., 2003). Separation specifies the distinct number of item or participant groups measured by the latent variable (acceptable ≥ 1.50, good ≥ 2.00, excellent ≥ 3.00; Duncan et al., 2003). Average standard error of measurement (SEM) for items provides a measure of test precision with lower values indicating greater measurement accuracy. Item infit and outfit mean-square statistics between 0.50 and 1.50 logits are most productive for measurement and anything greater than 2.00 could distort measurement (Linacre, 2002). Item point-biserial correlations must be positive in value to demonstrate they offer measurement support, while negative point-biserial correlations suggest item removal is necessary as these items contribute in opposition to the latent variable's meaning (Wright, 1992). With Rasch measurement, each item produces a difficulty measure in logits with higher values indicating an item is more challenging to answer correctly and lower values meaning an item is easier for students to correctly answer. Item difficulty measures were compared between administration formats and considered statistically similar if they fell within ±2 standard deviations.

Findings

In terms of overall test and item comparability between testing formats, all psychometrics indices were nearly identical and told the same story (see Table 2).

Table 2

Test and Item Performance Comparison

Test and Item Psychometrics	Testing Format	
	Paper-Pencil	Computer-Based
Item Reliability	0.91	0.93
Item Separation	3.63	3.57
Average Standard Error	0.62	0.76
Negative Point-Biserial Items	0 (0%)	0 (0%)
Misfitting Items	1 (2.3%)	1 (2.3%)
Statistically Easier Items	4 (9%)	2 (4.5%)

To summarize: Item reliability and separation were "Excellent" for both (Paper-Pencil = 0.91, Computer-Based = 0.93), SEM was approximately the same (Paper-Pencil = 0.62, Computer-Based = 0.76), no items had negative point-biserial correlations, and only one item was misfitting in each version. In terms of item difficulty: Among the 44 items across the 11 tests compared, 38 items (86%) performed statistically similarly (within ±2 standard deviations) regardless of the testing format. Items that differed statistically in their difficulty measure were

relatively balanced with similarly small numbers being easier when delivered in paper-pencil (n = 4, 9%) and computer-based formats (n = 2, 4.5%).

Discussion and Next Steps

Our goal was to compare the consistency of test performance and the capacity of those items to measure student ability when delivered via paper-pencil or computer-based methods. While such comparisons for multiple-choice items have been widely presented in previous research (e.g., Puhan et al., 2007), a relatively small number of studies have explored constructed response options (e.g., McClelland & Cuevas, 2020), and even fewer comparisons have been made using mathematical problem-solving tests (e.g., Lemmo & Mariotti, 2017). Two notable findings were observed in our comparison study. First, no significant or practical differences were observed relative to overall test performance (e.g., Rasch reliability, separation, item statistics) when implementing in either delivery format. Second, the overall capacity for PSM items to measure persons remains largely unchanged by delivery method.

Results from this study strengthen the evidence for using PSMs and comparing results regardless of delivery mode (paper-pencil vs. static computer-based). Next steps in our work are to test the computer-based items in a computer adaptive testing (CAT) delivery mode, as part of the design science approach. Our ability to compare the outcomes associated with delivery models *before* adding the CAT component helps to ensure that any differences uncovered during this phase are not simply the result of a change in delivery format. Given that results from delivery format comparisons have widely varied (Hamhuis et al., 2020), it is critically important that anyone considering moving a test from paper-pencil to computer-based delivery build in time to test comparability of overall assessment and item psychometrics.

Acknowledgement

Ideas in this manuscript stem from grant-funded research by the National Science Foundation (NSF 2101026, 2100988). Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation.

References

Bostic, J. D., & Sondergeld, T. A. (2015). Measuring sixth-grade students' problem-solving: Validating an instrument addressing the mathematics common core. *School Science and Mathematics Journal, 115*(6), 281-291.

Bostic, J. D., Sondergeld, T. A., Folger, T., & Kruse, L. (2017). PSM7 and PSM8: Validating two problem-solving measures. *Journal of Applied Measurement, 18*(2), 1-12.

Bostic, J., May, T., Matney, G., Koskey, K., Stone, G., & Folger, T. (2024, March). Computer

adaptive mathematical problem-solving measure: A brief validation report. In D. Kombe & A. Wheeler (Eds.), *Proceedings of the 51st Annual Meeting of the Research Council on Mathematics Learning* (pp. 102-110). Columbia, SC.

Common Core State Standards Initiative. (2011). *Common Core State Standards for Mathematics*. http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

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

Hamhuis, E., Glas, C., & Meelissen, M. (2020). Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students? *British Journal of Educational Technology*, *51*(6).

Keldermen, H. (1988). Common item equating using the loglinear Rasch model. *Journal of Educational Studies*, *13*(4), 319-336.

Lemmo, A., & Mariotti, M. A. (2017, February). From paper and pencil-to computer-based assessment: Some issues raised in the comparison. In *CERME 10*.

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), 878.

McClelland, T., & Cuevas, J. (2020). A comparison of computer based testing and paper and pencil testing in mathematics assessment. *The Online Journal of New Horizons in Education*, *10*(2), 78-89.

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.

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

Puhan, G., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper & Pencil & Computerized Testing. *Journal of Technology, Learning, & Assessment.*

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), with foreward and afterword by B.D. Wright. The University of Chicago Press.

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

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*, *66*, 335-348.

VanDerHeyden, A. M., Codding, R., & Solomon, B. G. (2023). Reliability of computer-based CBMs versus paper/pencil administration for fact and complex operations in mathematics. *Remedial and Special Education*, *44*(2), 91-101.

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. (1992). Point-biserial correlations and item fits. *Rasch Measurement Transactions, 5*(4), 174.