Received: 6 November 2020 | Revised: 18 January 2023 | Accepted: 23 January 2023

DOI: 10.1111/ssm.12570

RESEARCH PAPER – MATHEMATICS EDUCATION

WILEY | SSMA

Examining how using dichotomous and partial credit scoring models influence sixth-grade mathematical problem-solving assessment outcomes

Toni A. May[1] | **Kristin L. K. Koskey**[1] | **Jonathan D. Bostic**[2] | **Gregory E. Stone**[3] | **Lance M. Kruse**[4] | **Gabriel Matney**[2]

[1]Drexel University, School of Education, Philadelphia, Pennsylvania, USA

[2]Bowling Green State University, College of Education, Bowling Green, Ohio, USA

[3]University of Toledo, College of Education, Toledo, Ohio, USA

[4]Relias, Morrisville, North Carolina, USA

Correspondence
Toni A. May, Drexel University, School of Education, 3401 Market St., Philadelphia, PA 19104, USA.
Email: tas365@drexel.edu

Funding information
National Science Foundation, Grant/Award Numbers: NSF#1720646, 1720661, 2100988, 2101026

Abstract

Determining the most appropriate method of scoring an assessment is based on multiple factors, including the intended use of results, the assessment's purpose, and time constraints. Both the dichotomous and partial credit models have their advantages, yet direct comparisons of assessment outcomes from each method are not typical with constructed response items. The present study compared the impact of both scoring methods on the internal structure and consequential validity of a middle-grades problem-solving assessment called the problem solving measure for grade six (PSM6). After being scored both ways, Rasch dichotomous and partial credit analyses indicated similarly strong psychometric findings across models. Student outcome measures on the PSM6, scored both dichotomously and with partial credit, demonstrated strong, positive, significant correlation. Similar demographic patterns were noted regardless of scoring method. Both scoring methods produced similar results, suggesting that either would be appropriate to use with the PSM6.

KEYWORDS

assessment, constructed response items, dichotomous scoring, partial credit scoring

1 | INTRODUCTION

While there are a wide variety of options available for assessing students' mathematical abilities, the vast majority collect data in one of two ways: selected- (e.g., multiple choice, true/false, matching) or constructed-response items (e.g., word problems, authentic, performance assessments) (Brookhart & Nitko, 2019; McMillan, 2011; Mertler, 2003; Popham, 2014). Mathematics assessments where students select a response provide students with an item and then give answer options for students to choose from. In comparison, constructed-response items require students to develop (or construct) their own mathematical answer to an item (e.g., solve the problem, fill-in-the-blank, lab report). Dichotomous and partial credit scoring are two main methods for grading assessments. *Dichotomous scoring* assesses student responses as correct or incorrect (usually 1 or 0), whereas *partial credit scoring* reflects levels of correctness and allows for students to receive a range of scores based on correctness of response (Bond & Fox, 2007).

Specifying which scoring procedure is used is a necessary component of assessment development and validation to yield highly consistent and meaningful test scores for use (American Educational Research Association [AERA] et al., 2014). The Problem-Solving Measure for grade 6 (PSM6) is a measure of mathematical problem-solving consisting of constructed-response items. Important in test development is to systematically examine

how different scoring methods compare to adopt the method yielding the most accurate and meaningful scores (Jiao et al., 2012). Although multiple sources of validity and reliability evidences have been published on the PSM6, (Bostic & Sondergeld, 2015), investigation of the measurement properties using the current dichotomous scoring as compared to other methods continues to be needed.

As such, this study compared the impact of two scoring methods on internal structure and consequential sources of validity evidence for the PSM6. Findings aim to contribute to a multi-year validity study on the PSM6 (Bostic & Sondergeld, 2015) and be utilized to advance the PSM6 for wider educational use. Research is limited in how dichotomous and partial credit scoring compare when applied to constructed-response items. Thus, this study is also particularly relevant to the field of mathematical education at-large because tests of student mathematical abilities are often assessed through open-ended, constructed-response item types to evaluate higher level thinking skills such as problem solving (see Bostic & Sondergeld, 2015).

2 | BACKGROUND

2.1 | Problem solving in mathematics education

In the present study, *mathematical problem-solving* is defined similarly to Lesh and Zawojewski (2007) as "the process of interpreting a situation mathematically, which usually involved several cycles of expressing, testing, and revising mathematical interpretations" (p. 782). Problem solving requires students to solve problems (Kilpatrick et al., 2001), not exercises. Mathematical exercises are tasks intended to advance student content proficiency through a known procedure (Kilpatrick et al., 2001). According to Schoenfeld (2011), students engage in problem solving when mathematical tasks (i.e., problems) include three elements: (a) solutions, (b) numerous pathways, through which a solution for the task could be sought; and (c) multiple potentially correct responses. It is possible that "no solution" is a mathematical result to a given task; thus, meeting the solution element of a problem. Furthermore, mathematical problem-solving tasks should be developmentally appropriate for students while possessing characteristics of openness (multiple solution methods), realism (draw on experiential knowledge), and complexity (sustained reasoning required to solve) (Boaler & Staples, 2008; Palm, 2006; Verschaffel et al., 1999).

Globally, teaching primary and secondary students mathematical problem-solving has emerged as a prominent theme threaded throughout educational standards to various degrees (Common Core State Standards Initiative [CCSSI], 2010; Mullis et al., 2016; National Council of Teachers of Mathematics, 2000, 2014). Approximately two-thirds of the 56 countries participating in the international *Trends in Mathematics and Science Study* (*TIMSS*) reported specific mathematical problem-solving standards for their country's students (Mullis et al., 2016). Countries that explicitly address mathematical problem solving in their primary and/or secondary education include but are not limited to: Australia, Canada, Chinese Taipei, Cyprus, Denmark, Finland, France, Germany, Ireland, Japan, Singapore, Spain, Sweden, Thailand, Turkey, and the United States (U.S.) (Mullis et al., 2016). For a more comprehensive view of mathematical standards by country, readers might consult the *TIMSS* 2015 *Encyclopedia* (http://timssandpirls.bc.edu/timss2015/encyclopedia/countries/#side).

2.2 | Measures of mathematical problem-solving skills

Four categories of measures of mathematical problem-solving skills have been developed to date (Bostic et al., 2022). Category 1 includes large-scale standardized tests of mathematical problem-solving including the National Assessment of Educational Progress (NAEP) and Programme for International Student Assessment (PISA). Category 2 consists of tests developed from a psychological perspective such as the mathematics problem solving subtest included on the Woodcock-Johnson IV Tests of Achievement (Schrank & Wendling, 2018) and Wechsler Intelligence Scale for Children (Grizzle, 2011). Category 3 is made up of a number of unnamed assessments developed for specific research studies (e.g., Charles & Lester, 1984; Verschaffel et al., 1999). Category 4 comprises tests for the purpose of progress monitoring such as the Measure of Academic Progress (Meyer & Dahlin, 2022), iXL (Bashkov et al., 2021), and STAR Math© (Renaissance Learning, 2022) used at-large in K-12 schools across the U.S.

Problem-solving measures (PSMs, Bostic & Sondergeld, 2015; Bostic et al., 2017) are a series of measures developed for grades 3–8 that contribute to the advancement of the assessment of mathematical problem-solving skills beyond these four categories of measures in two ways. First, PSMs are distinct in that they are fully aligned with the Common Core State Standards for Mathematics (CCSSM;, 2010) which are currently implemented across 41 states in the U.S. Because problem solving is internationally recognized as mathematical practice that should be taught to primary and secondary students, it stands to reason that teachers

should implement assessments to measure their students' problem-solving growth and abilities in relation to curricular standards. Second, all PSM items require students to construct a response to a real-world scenario. Consistent with the definition of problem-solving adopted in this study, PSMs require students to apply "mathematical concepts from various topics and within and beyond mathematics" (Lesh & Zawojewski, 2007, p. 782).

2.3 | Problem-solving measure for grade 6

The PSM6 is part of the series of PSMs and composed of 15 items aligned with one or more standards for mathematics content (SMC) from the CCSSI (2010). PSM assessment results are designed to be used as both a formative and summative measure of students' mathematics content knowledge. Students are administered the PSM6 in a pre-post format to assess growth in learning from the beginning of year to end. As such, results of pre-tests are used to provide teachers with information on student strengths as well as areas of needed growth to better inform instruction over the course of the year. This is easily achieved because PSM6 items are directly aligned with grade level instructional standards. Additionally, end of year growth results help teachers see areas they may need to adjust instruction for in the following academic year with new students.

This instrument has undergone a multi-year rigorous validation study and demonstrated sufficient validity evidence on multiple indicators (Bostic & Sondergeld, 2015) aligned with *The Standards for Educational and Psychological Testing* (AERA et al., 2014). To summarize validity evidence results found from prior research, each item was evaluated by an expert panel consisting of mathematics teachers, mathematics educators, and mathematicians, who determined that the items were consistently worded to capture mathematical problem solving skills. Items were noted as complex, realistic, containing multiple solution pathways, and possessing a well-defined solution set, thereby supporting content validity evidence. Response process validity evidence was gathered through student think-aloud tasks, which revealed students were solving PSM6 items similarly to how item developers had hypothesized. When asked about how completing the PSM6 made students feel, students reported the assessment was challenging but felt no negative impact, suggesting consequential validity evidence was strong. Numerous psychometric indices (including Rasch item fit, reliability, separation, etc.) have all produced acceptable findings supporting internal structure validity evidence and high (0.97) internal consistency. Finally, relationship to other variables validity evidence has been evaluated by looking for differences in PSM6 findings by gender, race/ethnicity, and teachers' perceived ability level of their students. As hypothesized by the research team, there were no significant differences in PSM6 outcomes by gender or race/ethnicity, but there were by student ability level. Collectively, these validity evidences support the use of PSM6 for measuring mathematical problem-solving skills; however, additional validity evidence is needed related to the scoring method prior to wider-use in practice.

2.4 | Defining mathematical problem-solving scoring methods

Since the PSM6 consists of constructed-response items, dichotomous and partial credit scoring are two potential methods to compare for adoption. Dichotomous scoring involves scoring an item response as correct/incorrect. This method can be used for scoring selected-response or constructed-response type items focusing scoring on only the final solution. Partial credit involves scoring aspects of the problem-solving process shown and is more often used for scoring selected constructed-response type items (Brookhart & Nitko, 2019; Mertler, 2003). The two different scoring methods might be implemented in grading a single assessment, such as when an assessment consists of both selected- and constructed-response items. Regardless of dichotomous or partial credit, a scoring key or rubric is necessary for producing highly reliable scores for use in educational practice and research (Jonsson & Svingby, 2007; Mertler, 2001).

A scoring key specifying the correct answer(s) is used to guide assigning a dichotomous score. Scoring for only answer correctness has the very real and practical advantage of speed in grading (McMillan, 2011; Mertler, 2003). However, a fundamental disadvantage to dichotomous scoring lies in its all or nothing approach, such that the reasons a student arrived at an incorrect response were discounted (Lau & Wang, 1998; Rogers & Ndalichako, 2000). Such concerns with dichotomous scoring have led others to favor a partial credit approach (Grunert et al., 2013), which makes use of a set of criteria to determine the extent and potentially type of misunderstanding within incorrect student responses (Jiao et al., 2012) in a more direct fashion.

For partial credit scoring, a rubric is applied specifying levels of performance on different aspects of the mathematical problem solving (Brookhart & Nitko, 2019). Multiple decisions inform the rubric design, type of rubric utilized, and weighting of criteria (Brookhart & Nitko, 2019). Criteria are guided by the aspects of the mathematical

problem solving process to be assessed. A holistic rubric evaluating multiple criteria collectively to produce an overall rating or an analytic rubric evaluating each criterion separately can be adopted (Mertler, 2001). Further, a general rubric might be applied describing criteria and levels of performance in problem-solving skills consistently across all items on an assessment (Brookhart & Nitko, 2019). Alternatively, a task-specific rubric tailored to the content or process assessed on a given mathematical problem-solving item could be implemented (Brookhart & Nitko, 2019). Criteria might be assigned equal or varying weight using different rating scales, points, or weighted percentages. Multiple factors can be considered in assigning weight such as whether a criterion is linked to the primary learning outcome, degree of precision, or product/final solution.

2.5 | Research comparing dichotomous and partial credit scoring

Literature on how to use various scoring methods or which methods are most appropriate for varying assessment purposes, is plentiful. However, research directly comparing dichotomous and partial credit scoring on the same constructed-response type items is sparse. The studies that do exist, predominantly focus on multiple-choice tests. This should not come as a surprise since "as long as multiple choice tests have been in general use, there seems to have been a widespread, nagging uneasiness about the all-or-nothing character of conventional number-right scoring" (Frary, 1989, p. 79). As a result, studies have worked to find reliable and valid methods for providing partial credit for student knowledge on these widely used assessments, while neglecting other item types.

Though there are many methods of comparing dichotomous and partial credit scoring with multiple choice items (see Frary, 1989), much of the work in this area uses multiple-select multiple-choice (MSMC) type items. MSMC items are also known as Pick-N or multiple true/false items because test-takers are asked to select as many response options as they believe are correct which turns each option into its own true/false item. The MSMC item also lends itself nicely to both scoring methods as partial credit can easily be awarded for demonstrating no knowledge (correctly selecting none of the true responses), partial knowledge (correctly selecting some of the true responses), or complete knowledge (correctly selecting all of the true responses). And dichotomously scored MSMC items are either correct (all true responses selected) or incorrect (anything less than all correct).

Frary (1989) conducted a comprehensive review of research comparing partial credit and dichotomous scoring for multiple choice tests. He concluded that while some partial credit methods increased internal consistency reliability, there was not enough evidence from a single study or the collective group that suggested it was a better scoring option in comparison to dichotomous scoring. This suggestion arose because there were too many other factors that devalued the use of partial credit: "reduced validity, increased time required for testing, scoring complexity, difficulty of explaining the scoring to examinees and other users, and difficulty of explaining the response mode and training examinees in its use" (Frary, 1989, p. 92). Frary further believed that in the few instances where partial credit was considered appropriate, it was due to desirable derivatives such as providing feedback to test takers, and not to any supposed psychometric benefits.

Since Frary's (1989) research synthesis, studies have been conducted on the same topic with varying results related to item discrimination, student classification, and reliability of estimates of student ability. In terms of item discrimination, partial credit has been shown to discriminate test taker ability better since dichotomously scored items are comparatively more difficult (Bauer et al., 2011; Ripkey et al., 1996), although the differences in item discrimination were relatively small. With relation to student ability or classification, theoretically, one would expect partial credit models to perform better since information is being provided over a larger ability range (Jiao et al., 2012). However, results have been mixed, with test taker classification decisions found to be equivalent regardless of the scoring method in some studies (Grunert et al., 2013; Jiao et al., 2012) or producing slightly higher estimates of student ability with simulated data (Jiao et al., 2012).

Similar to Frary's (1989) conclusions, other studies have found partial credit scoring of multiple choice items to elicit higher levels of reliability compared to dichotomously scored items (Albanese & Sabers, 1998; Bauer et al., 2011; Ripkey et al., 1996). This is in part because dichotomous scoring offers less information about student ability by classifying a student who correctly answers two out of three options the same as a student who correctly answers none. Further, partial credit scoring has been shown to result in higher item total correlations (Ripkey et al., 1996). This finding is also theoretically sound because reliability is a correlational measure which depends on variability; the greater variability that exists from a wider range of partial credit scores allows for a stronger likelihood of item correlation.

After an extensive review of the literature, we were unable to find any studies that compared dichotomous

and partial credit scoring on the same set of constructed-response items. This may in part be because some believe that "a scorer would probably be considered inflexible or possibly negligent" (Frary, 1989) if they failed to use partial credit scoring for constructed response items. Nevertheless, there are times when either scoring method may be appropriate. For example, a mathematics problem-solving item, where students are asked to produce their own response, may be scored dichotomously if the purpose is to assess only the answer, or it may be scored with partial credit if both the process (student work) and product (final solution) were to be assessed. The current study contributes not only to the advancement of sources of validity evidence for the PSM6, but also to informing the limited literature on the differential effects of the two methods on students' scores for constructed-response items.

2.6 | Purpose and research questions

The primary purpose of this study was to evaluate validity evidences for the PSM6 when using dichotomous and partial credit scoring. A secondary purpose of this study was to inform the larger body of literature on the differential effects of the two scoring methods for constructed-response items. AERA et al.'s (2014) *Standards for Educational and Psychological Testing* guided this study. Internal structure and consequences from testing and bias sources of validity evidence were examined specifically through three research questions:

1. Did the scoring method (dichotomous or partial credit) have a differential impact on the PSM6's estimates of item difficulty and student ability, unidimensionality, and item discrimination (internal structure)?
2. Was there a significant relationship between PSM6 scores when analyzed using dichotomous and partial credit scoring methods (internal structure)?
3. Did the scoring method (dichotomous or partial credit) have a differential impact on detecting systematic and observed gender group differences on the PSM6 (consequential/bias)?

3 | METHODS

3.1 | Sample

Teachers administered the assessment to all students in their mathematics classes in the last month of school to ensure students had been given an opportunity to be introduced to all mathematical content covered on the PSM6. Data were collected from 517 sixth-grade students across eight Midwest U.S. schools and 16 classrooms. Student gender was teacher-reported. No other student demographic data were collected for this sample. A total of 261 (50.5%) boys and 243 (47.2%) girls completed the PSM6 (gender was not reported for 12 or 2.3% of students).

3.2 | Scoring of the PSM6 items

Prior to analysis, grading was completed using both dichotomous and partial credit scoring on every item using two different generic rubrics for each scoring method. Three evaluators, each having earned a state license to teach mathematics, were trained on scoring procedures with a subset of PSM6 assessments by rating and discussing their scoring. Interrater agreement was extremely high (1.00) because all evaluators had been working on this project for multiple years and had been scoring PSM6 items as part of their project responsibilities. Once evaluators were certain they were scoring similarly, they divided the PSM6 assessments evenly to be scored independently. Evaluators were scoring PSM6 items for the purpose of this research study, and no other reason at the time. They first scored all items dichotomously. Dichotomous scoring was marked as correct (fully correct response = 1 point) or incorrect (partially or completely incorrect response = 0 points).

After scoring items dichotomously, evaluators went back through the incorrect responses and scored them for partial credit. Partial credit scoring was completed using the same holistic, general rubric applied across all PSM6 items. The rubric allowed for student responses to be scored as correct (fully correct response = 2 points), partially correct (incorrect answer but correct strategy [representation or procedure] = 1 point), or incorrect (incorrect answer and missing/incorrect strategy = 0 points). Scoring for partial correctness emerged from mathematics education literature on the practice of implementing partial credit scoring. It required consistent attention to both representations and procedures because the effective use of both is a key component of problem solving (Kilpatrick et al., 2001; Verschaffel et al., 2000).

Decisions about assigning partial credit were informed by past research (Bostic et al., 2011) and alignment with current frameworks. First, past peer-reviewed research used pilot versions of PSM items. Those items were scored using a three-point scale as used in the present study. For context, a mathematical strategy is defined as having both (a) mathematical representation

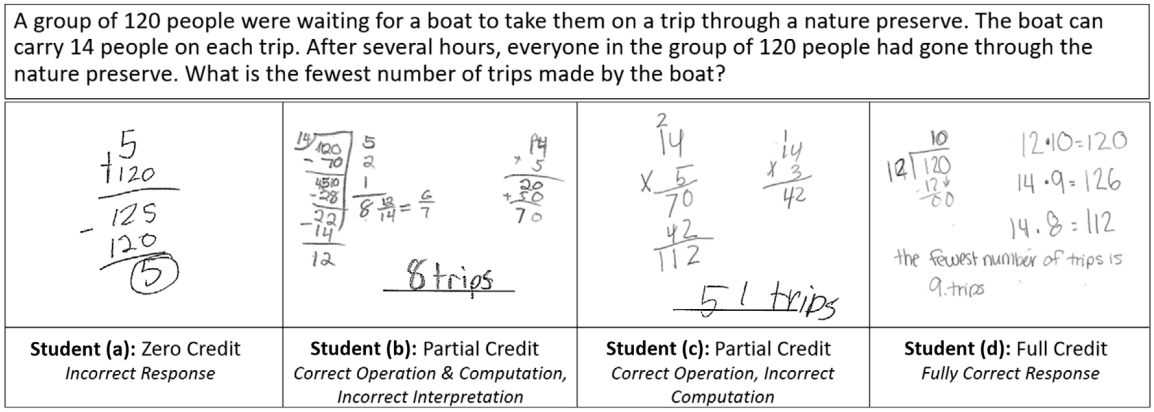

FIGURE 1 Sample PSM6 item with four examples of scored student responses.

and (b) mathematical procedure (Goldin, 2002; Verschaffel et al., 2000). An error in a mathematical representation, mathematical procedure, or interpreting the result from the strategy can potentially to lead an incorrect result for a given problem, which was seen in Bostic et al. (2011). Second, the scoring protocol for the present research was also influenced by past research that leveraged Verschaffel et al.'s (2000) problem-solving framework that values mathematical work prior to reporting the solution. It has been observed that students apply a correct strategy but misinterpret their solution and report an incorrect answer (see Verschaffel et al., 2000). This prior research further supports the importance of exploring whether assigning partial credit is a more effective scoring method than dichotomous scoring. We drew these ideas together to define *operation* as representation and procedure usage. *Computation* was operationalized as carrying out arithmetic with the operation(s) correctly. Ultimately, our research team operationalized partial credit scoring as (a) correct operation usage and correct computation with incorrect interpretation of results or (b) correct operation usage and incorrect computation or vice versa.

Figure 1 presents work from four different example students on the same PSM6 item along with an explanation of scoring to demonstrate how our mathematical scoring practices were implemented. Items not attempted (i.e., no work shown on page) were considered missing data and were not scored as incorrect so that students were not penalized for not completing a problem as it could not be determined why the item was not attempted. Evaluators on this study were able to reliably dichotomously score approximately 60 PSM6 assessments per hour and 20 PSM6 assessments per hour when using partial credit to score items. Student overall ability measures on the PSM6 were estimated using the Rasch (1960)/(1980) model as described in the Data Analysis section.

3.3 | Data analysis

Rasch (1960)/(1980) measurement was used to inform RQ1 through Winsteps Version 4.4.06 (Linacre, 2012). The Rasch model was selected for this analysis because of its ability to examine the hierarchical linear scale produced by responses to an instrument, the dimensional and expectational foundations for the instrument, and the fullness of item performance. It was further selected because the model seamlessly allows for the adoption of either a dichotomous or partial credit analytic model. In both models, students' item level scores were transformed into to an overall person ability score in logits along the linear measure with a person mean set at 0 logits.

The Rasch model for dichotomous data (Wright & Stone, 1979) employs a maximum likelihood statistical estimation of student measures within the framework of correct and incorrect response data. A dichotomous model was applied and expressed in log-odds as

$$\ln(\pi_{ni1}|\pi_{ni0}) = \beta_n - \delta_i$$

where π (pi) is the response probability for student n, β (beta) is the student ability, and δ (delta) is the item difficulty for a given item i (Bond & Fox, 2007). The Rasch partial credit model (Masters, 1982) extends Andrich's rating scale model to allow for a hierarchical series of responses from completely incorrect to completely correct adhering to the same fundamental requirements necessary for the production of linear measures. In the Rasch partial credit model, threshold estimates can vary for each item as expressed below where k is the threshold for item i:

$$\ln(\pi_{nik}|1-\pi_{nik}) = \beta_n - \delta_{ik}$$

When applying either scoring model (dichotomous or partial credit), an overall ability measure is computed for

TABLE 1 Rasch separation and reliability statistics for dichotomous and partial credit scoring methods ($N = 515$).

	Person		Item	
Scoring method	**Separation**	**Reliability**	**Separation**	**Reliability**
Dichotomous	1.39	0.66	8.60	0.99
Partial credit	1.78	0.76	11.92	0.99

each student as a function of the student ability and item difficulty. This ability measure is transformed into log-odd units (i.e., logits) along the linear PSM6. Resultant student logit ability measures from the psychometric analyses were used to conduct traditional statistical analyses in SPSS descriptively and with a Pearson Correlation to examine the relationship between the scoring method and students' measures on the PSM6 to inform RQ2.

Differences in detecting systematic DIF by gender sub-group was tested by conducting Rasch DIF analyses also using Winsteps Version 4.4.06 (Linacre, 2012) to inform RQ3. Items yielding a $t \geq \pm 1.96$ ($n = 505$; $df = 503$, $\alpha = 0.05$) indicated statistically significant DIF. Differences in item difficulty between boys and girls were computed using DIF contrast sizes. According to Zwick et al. (1999), DIF contrast sizes $\geq$ 0.43 logits indicate moderate to large magnitude in differences. Independent samples *t*-tests were conducted in SPSS to further test for differential influence on detecting observed differences by gender group using student logit ability measures from dichotomous and partial credit psychometric analyses.

4 | FINDINGS

4.1 | Scoring model impact on PSM6 item and person outcomes (RQ1)

4.1.1 | Reliability

Rasch reliability is similar to traditional reliability as both assess internal consistency. Computationally, because of the inclusion or exclusion of extreme data, Rasch reliability tends to underestimate reliability, while traditional Cronbach alpha tends to overestimate it. Therefore, using the more conservative Rasch indicator was considered appropriate. Separation indicates the number of statistically distinguishable groups that can be classified on a variable. As such, separation can be seen as a measure of clarity. Both Rasch reliability and separation are respectively considered acceptable at 0.70 and 1.50; good at 0.80 and 2.00; and excellent at 0.90 and 3.00 (Duncan et al., 2003). Table 1 presents reliability and separation statistics for both dichotomous and partial credit models. Person separation and reliability were acceptable or nearly acceptable for both models, though marginally higher for the partial credit analysis. Item separation and reliability were considered excellent for both models, yet slightly higher for partial credit.

4.1.2 | Unidimensionality

There is no one best way to assess unidimensionality of a measure, thus multiple indicators are investigated. Items with negative point-biserial correlations or infit/outfit mean square (MNSQ) fit statistics falling outside 0.5–1.5 logits are not meaningful for measurement (Linacre, 2002). Table 2 reports item level MNSQ fit statistics, difficulty measures, and point-biserials. Item indices across the two scoring models are listed by item number with the aligned SMC domain indicated in Table 2.

For both dichotomous and partial credit runs, no PSM6 items had negative point-biserials and item infit was within meaningful parameters. Item 14 yielded a MNSQ outfit slightly below the 0.50 criterion applying the dichotomous (MNSQ = 0.28) and partial credit (MNSQ = 0.20) suggesting overfitting to the Rasch model. Both models resulted in one item with an MNSQ above the 1.50 criterion, indicating a lower degree of predictability in the responses. Outfit is more sensitive to extreme responses and one item slightly underfitting the Rasch model is not degrading to the measure (Linacre, 2002). Overall, PSM6 item fit was comparable when using both scoring models and supported undimensionality of the linear measure.

Rasch principal components analysis (RPCA) represents a second way to examine dimensionality. RPCA attempts to extract the common variance that best explains the residual variance (Linacre, 1998). The instrument, and the variable it attempts to measure, are considered theoretically unidimensional if at least 60% of the variance is explained. This is often difficult to achieve in content areas that are, in fact, generally unidimensional (e.g., arithmetic) yet include several strong content areas that can also function independently (e.g., addition, subtraction, and multiplication within arithmetic). Such was the case with the dichotomous model (44% variance explained) and the partial credit model (50% variance explained). Because neither model met the 60% variance explained criteria, it was important to further explore items within the first contrast. To achieve practical unidimensionality, less than 5% of the variance should be explained by the first

TABLE 2 PSM6 standards for mathematics content alignment and item statistics for dichotomous and partial credit scoring methods ($N = 515$).

Dichotomous scoring								Partial credit scoring					
Difficulty in logits	***SE***	**MnSQ infit**	**MnSQ outfit**	***pb***	**DIF contrast**	**Item**	**SMC domains primary/secondary**	**Difficulty in logits**	***SE***	**MnSQ infit**	**MnSQ outfit**	***pb***	**DIF contrast**
−2.97	0.12	0.98	1.08	0.70	0.28	1	6.SP.1	−1.68	0.07	0.87	0.84	0.72	0.20
3.51	0.47	0.91	0.64	0.20	−0.38	2	6.G.1	2.01	0.15	0.98	1.04	0.30	−0.07
−0.93	0.13	1.14	1.39	0.50	−0.40	3	6.NS.3/NS.1	−0.34	0.07	1.19	1.49	0.52	−0.22
−1.88	0.12	0.94	1.00	0.70	−0.22	4	6.RP.3/EE.7	−1.50	0.07	1.00	1.00	0.67	−0.13
−0.39	0.14	1.08	1.42	0.50	−0.64*	5	6.NS.3	−0.07	0.07	1.17	1.68	0.48	−0.36*
−1.02	0.13	1.10	1.25	0.50	0.10	6	6.EE.7	−0.50	0.07	1.11	1.30	0.55	0.00
2.83	0.36	1.05	2.20	0.20	0.61	7	6.G.2	1.72	0.15	1.03	1.03	0.27	0.11
−2.83	0.13	0.96	0.86	0.70	0.00	8	6.SP.1	−1.55	0.07	0.94	0.84	0.70	0.00
−1.35	0.13	0.91	0.82	0.70	0.00	9	6.EE.2	−0.60	0.07	0.93	0.80	0.62	0.00
0.86	0.19	0.98	0.99	0.40	1.19**	10	6.RP.3	0.72	0.10	1.03	1.05	0.40	0.52*
−0.19	0.15	0.95	0.90	0.50	0.00	11	6.NS.3	0.07	0.08	0.98	0.92	0.53	0.00
−0.04	0.16	0.93	0.77	0.50	0.00	12	6.RP.3	−0.14	0.08	0.91	0.83	0.60	0.14
2.66	0.34	1.04	0.64	0.20	0.47	13	6.SP.5	1.13	0.11	0.96	0.98	0.50	0.07
3.25	0.43	0.93	0.28	0.20	0.00	14	6.G.4/G.1	2.01	0.19	0.85	0.20	0.29	−0.10
−1.69	0.13	0.94	0.95	0.70	0.14	15	6.RP.3/EE.7	−1.26	0.08	0.97	0.97	0.67	0.15

Note: MnSQ outside 0.5–1.5 logits not meaningful for measurement (Linacre, 2002).

Abbreviations: DIF, differential item functioning; EE, expressions and equations; G, geometry; MnSQ, mean square; NS, number sense; RP, ratio and proportions; SMC, standards for mathematics content domain; SP, statistics and probability.

*$p < 0.05$; **$p < 0.01$.

TABLE 3 Person ability mean statistics in logits and item discrimination information for dichotomous and partial credit scoring models ($N = 515$).

	Person ability			Item discrimination		
Scoring method	*M*	*SD*	Statistically similar range	Too easy (poorly discriminates)	Similar (discriminates well)	Too difficult (poorly discriminates)
Dichotomous	−1.96	1.46	0.96 to −4.88	0 items	11 items	4 items
Partial credit	−1.12	1.08	1.04 to −3.28	0 items	11 items	4 items

contrast of items. Both dichotomous (6.0%) and partial credit models (5.5%) demonstrated this approximate level of statistical performance. Additionally, when considering the question of unidimensionality, if the first contrast three or more items, then a meaningful contrast of items may exist within the measure, detracting from its overall status. Within both scoring models, only two items loaded at 0.40 or higher, suggesting unidimensionality regardless of scoring model. Thus, although neither dichotomous nor partial credit scoring model RPCAs reached the desired goal of explaining 60% of the variance, it is important to note that no meaningful alternative dimensions were found.

4.1.3 | Item discrimination

Items that discriminate well, can categorize students into various ability groups within a specified margin of error. Items that are either too easy or too difficult for students do not discriminate amongst student ability groups because they are considered extreme. To evaluate the item discrimination within each scoring model, the person mean and standard deviation from each of the models were calculated in order to define a statistical confidence interval. A confidence interval approach to item discrimination was selected in order to make comparisons of the measures more feasible where anchoring is not possible. Two times the person mean standard deviation was both added and subtracted from respective person means to generate a normal range that item difficulties would need to fall within to be considered statistically similar to their corresponding person mean ability. Items that fell outside of this range in terms of difficulty measure were considered significantly different—either significantly difficult or significantly easy in comparison to the mean student ability.

Table 3 shows person mean ability statistics and number of significantly similar or different items by scoring method. For the 15 items on the PSM6, both scoring methods produced the same number of significantly difficult items ($n = 4$ items, 26.7%), no significantly easy items, and a majority of items falling within a statistically similar range ($n = 11$ items, 73.3%).

Item difficulty (in logits) were reported in Table 2 for both models. The distribution of item difficulty in relation to person ability along the linear measure were plotted side-by-side for the dichotomous and partial credit models in Figure 2.

For both models, item ordering was similar along the linear measure. Also, over a majority (>73%) of the items yielded an item difficulty above the person mean ability measure, suggesting better targeting of students with higher mathematical problem-solving ability for the dichotomous and partial credit scoring models. This finding supports the continued inclusion of linking items from the PSM5 (grade level below) to assess student ability at the lower end of the continuum of mathematical problem-solving.

4.2 | PSM6 student measure relationship by scoring model (RQ2)

There was a statistically significant, very high, positive relationship between student PSM6 measures when scored dichotomously or by partial credit methods; $r(515) = 0.917$, $p < 0.001$, two-tailed. The effect size was very high ($r^2 = 0.84$) with 84% of the association in PSM6 scores accounted for by scoring method. To investigate this pattern more closely, students were divided into quartiles by their ability measures from each scoring method. After adding and subtracting corresponding standard error of measurement (*SEM*) to student ability measures, 99.61% ($n = 515$) were classified in the same quartile regardless of scoring method. The two students (0.39%) who were not classified in similar quartiles were placed in the 3rd quartile for partial credit scoring and the 1st quartile when scored dichotomously.

4.3 | Examining gender group differences (RQ3)

DIF results showed that both models detected DIF for item 5 and item 10. DIF contrasts were reported in Table 2. Item 5 DIF contrasts were a negative value when

PSM6 Item-Person Ordering for Dichotomous and Partial Credit Scoring Methods (N = 515)

Dichotomous Scoring

Person Measures	Item Measures
4 +	
Higher Score	**Higher Difficulty**
|	2 (6.G.1)
|	14 (6.G.4/G.1)
|	
3 +	7 (6.G.2)
|	
|	13 (6.SP.5)
. |	
|	
|S	
2 +	
. |	
|	
|	
|	
.## |	
1 T+	
|	10 (6.RP.3)
|	
.# |	
|	
. |	
0 .### +***M***	12 (6.RP.3)
. |	11 (6.NS.3)
. |	5 (6.NS.3)
.###### S|	
. |	
. |	
-1 .#### +	3 (6.NS.3/NS.1) 6 (6.EE.7)
|	
####### |	9 (6.EE.2)
.# |	
. |	15 (6.RP.3/EE.7)
.###### |	4 (6.RP.3/EE.7)
-2 . ***M***+	
.# | S	
.######### |	
. |	
. |	
.# |	1 (6.SP.1) 8 (6.SP.1)
-3 .######### +	
|	
Lower Score S|	**Lower Difficulty**
.# |	
.############## |	
|	
-4 +	

Partial Credit Scoring

Person Measures	Item Measures
4 +	
Higher Score	**Higher Difficulty**
|	
|	
|	
3 +	
|	
|	
|T	
|	
|	
2 . +	14 (6.G.4/G.1) 2 (6.G.1)
. |	7 (6.G.2)
|	
. |	
. |	
|S	13 (6.SP.5)
1 # T+	
.# |	
.# |	10 (6.RP.3)
.# |	
. |	
.### |	
0 .### S+***M***	11 (6.NS.3) 5(6.NS.3)
###### |	12 (6.RP.3)
.### |	3 (6.NS.3/NS.1)
.### |	6 (6.EE.7)
#### |	9 (6.EE.2)
.##### |	
-1 .###### +	
##### ***M***|S	15 (6.RP.3/EE.7)
.#### |	
##### |	4 (6.RP.3/EE.7)
.###### |	1 (6.SP.1) 8 (6.SP.1)
.## |	
-2 .#### +	
.#### S|	
.# |	
# |T	
.##### |	
|	
-3 . +	
.# |	
Lower Score .### T|	**Lower Difficulty**
|	
|	
|	
-4 +	

Note. "#" = 5 students and "." = 1 to 4 students. ***M*** = mean person ability or mean item difficulty.

FIGURE 2 PSM6 Item-person ordering for dichotomous and partial credit scoring methods ($N = 515$). "#" = 5 students and "." = 1–4 students. M = mean person ability or mean item difficulty.

using both scoring models, indicating that the item was more difficult for equal ability girls than boys. Using the dichotomous scoring model, the item 5 difficulty for boys was −0.68 logits ($SE = 0.19$), while the item difficulty for girls was −0.04 ($SE = 0.22$). Similarly, using the partial credit scoring model, the item difficulty for boys was −0.22 logits ($SE = 0.10$) and 0.13 ($SE = 0.11$) for girls. Item 10 DIF contrast was a positive value when using both scoring models, indicating the item was more difficult for equal ability boys than girls. Using the dichotomous

TABLE 4 Independent samples *t*-test results for student PSM6 scores by gender for dichotomous and partial credit scoring models ($n = 503$).

	Gender				
Scoring method	**Girls *M*(*SD*)**	**Boys *M*(*SD*)**	***t*-statistic**	***p*-value**	**η^2**
Dichotomous	−2.67(1.89)	−2.72(1.83)	0.31	0.755	0.001
Partial credit	−1.32(1.34)	−1.55(1.48)	1.83	0.068	0.006

scoring model, the item 10 difficulty for boys was 1.56 logits ($SE = 0.34$), while the item difficulty for girls was 0.36 ($SE = 0.24$). When implementing the partial credit scoring model, the item difficulty for boys was 0.99 logits ($SE = 0.16$) and 0.47 ($SE = 0.13$) for girls. All other PSM6 items had no statistically significant DIF between boys and girls. Independent samples *t*-tests using overall student ability measures revealed no statistically significant observed differences by gender regardless of the scoring method ($p > 0.05$) (see Table 4). These findings suggest that overall, both scoring models were comparable in detecting systematic and observed differences by gender group.

5 | DISCUSSION

The Standards for Educational and Psychological Testing (AERA et al., 2014) specify the importance of determining a consistent scoring procedure to apply across examinees as an "integrated" process (p. 79). Thus, determining the most appropriate scoring method to implement to obtain student outcomes with the greatest reliability and validity evidence must be based on multiple facets, including appropriate alignment with academic standards, curriculum, and instruction (Brookhart & Nitko, 2008; Mertler, 2003), purpose for using assessment results (Brookhart & Nitko, 2008; Popham, 2014), as well as the practicality and efficiency of scoring model used (McMillan, 2011). While many studies have investigated the impact of applying dichotomous and partial credit scoring on assessments, this body of literature focuses on multiple-select multiple-choice item assessments (e.g., Bauer et al., 2011; Frary, 1989; Grunert et al., 2013; Jiao et al., 2012; Ripkey et al., 1996) because it is easy to apply both scoring models with such items. The current research explored possible advantages and disadvantages associated with the use of dichotomous and partial credit scoring models when evaluating sixth grade students' mathematical problem-solving ability on a constructed-response assessment (PSM6). Overall, our research findings provided sources of validity evidence supporting both dichotomous and partial credit scoring methods as yielding comparable measurement properties for constructed-response items on the PSM6. Results from this study demonstrated the importance of examining the effects from different scoring procedures prior to adoption of a scoring method (AERA et al., 2014; Jiao et al., 2012). Also, comparable results for the two scoring methods lends support for adopting the less costly and time intensive dichotomous scoring method for the PSM6 (McMillan, 2011).

5.1 | Comparing the two scoring models

5.1.1 | Psychometric comparisons

To evaluate the differential impact of the scoring model (dichotomous or partial credit), several indices were inspected. Rasch reliability and separation, along with measures of dimensionality were functionally equivalent. The slightly higher performance of the partial credit model is most likely due to the presence of more refined data (i.e., three rating scale points rather than two) (Bendig, 1954). Our reliability results are similar to those found in scoring method studies assessing MSMC item tests (Albanese & Sabers, 1998;Bauer et al., 2011; Ripkey et al., 1996). In our study, item discrimination values were identical regardless of scoring method as each identified the same number of items that were considered too easy, too difficult, and most appropriate for test takers. Taken collectively, our findings differed from earlier studies. In prior evaluations, assessing MSMC items, dichotomously scored tests most often identified slightly more difficult items compared to tests scored with partial credit (Bauer et al., 2011; Ripkey et al., 1996). Thus, the psychometric benefits and drawbacks of scoring model used to assess PSM6 constructed-response items indicate there may be some minor theoretical differences, but in practice these two methods were demonstrated to be quite similar psychometrically.

5.1.2 | Student measure comparisons

While one might hypothesize partial credit scoring would produce higher scores for students as ability information is given over a larger range (Jiao et al., 2012), past study findings from MSMC studies have been mixed (Grunert et al., 2013; Jiao et al., 2012). Our study showed that student performance of problem-solving content is nearly identical regardless of scoring

method chosen. In practical terms, students who had the highest or lowest PSM6 measures when scored dichotomously also had the highest or lowest measures when scored with the partial credit model, which is not surprising as student ability should be similarly representative of their content mastery regardless of scoring method. Further, nearly all students (99.61%) were placed in the same quartile irrespective of the scoring method. For the two anomalies (0.39% of students) in our dataset, findings did present in the direction literature would expect as these students were placed in the 3rd quartile for partial credit scoring and 1st quartile when scored dichotomously. Our student ability outcome comparison suggests there is no meaningful difference found by scoring method.

5.1.3 | Gender group comparisons

Both dichotomously scored and partial credit scored PSM6 assessments had comparable results in detecting observed and systematic differences in scores by gender group. Variance in test scores due to construct-irrelevant extraneous factors (e.g., gender group, item bias, scoring method) can distort the interpretation of the scores for specific sub-groups, negatively impacting consequential validity evidence (AERA et al., 2014). The current study results support consequential validity evidence of the PSM6 in two ways when using either scoring method. First, no significant sub-group differences were observed by gender regardless of scoring method. Second, less than 10% of the items exhibited DIF for the dichotomously scored and partial credit scored PSM6. Valid score interpretations across subgroups for the scoring method adopted is an essential component of consequential validity evidence and fairness in testing (AERA et al., 2014). Limited evidence of DIF is important to continue to support use of the PSM6 across a wider student population in the U.S. (AERA et al., 2014).

5.2 | Tentative conclusions and implications for practice

Many may feel uncomfortable using a dichotomous scoring model to evaluate constructed response items in general (Frary, 1989). Those assessing problem-solving performance in mathematics as student process view the importance of data and use partial credit scoring (e.g., Verschaffel et al., 1999). However, our study demonstrates empirical findings of near equivalence between partial credit and dichotomous scoring of PSM6 assessments across psychometric properties and student problem-solving ability classification outcomes. Such equivalence supports the use of the more straightforward dichotomous model for assessing constructed response mathematical items on the PSM6 when the goal is to determine student ability measures. When results are equivalent, the time and resources saved in scoring dichotomously (60 PSM6s per hour) compared to scoring with partial credit (20 PSM6s per hour) provides a very practical motivation for using the dichotomous method (McMillan, 2011).

While using dichotomous scoring for constructed response PSM6 items elicited similar student outcomes and provided practical benefits to our research team, the current study yielded sources of validity evidence supporting either scoring method. Additionally, it is important to keep in mind that partial credit scoring may offer mathematics teachers other benefits not investigated in this study when scoring constructed response items within the classroom. Partial credit scoring is certainly useful in the mathematics classroom where learning and remediation are designed to take place as it can provide feedback to students in ways that dichotomous scoring may not be able to do (Brookhart & Nitko, 2008; Frary, 1989; McMillan, 2011; Mertler, 2003). Further, if the purpose of the assessment is to evaluate both product (score) and process (work) to identify student misconceptions or areas of strength and weakness, dichotomous scoring is likely to not be most appropriate (Mertler, 2003; Popham, 2014). However, in the case of the PSM6, because items are directly linked to the CCSSM, teachers are able to see which domains and specific standards students have strength or weakness in even if the simpler dichotomous scoring technique is applied allowing for formative instructional decisions to be made.

5.3 | Recommendations for future research

We cannot generalize our findings about using a dichotomous scoring method with constructed response items to all types of mathematical tasks. Instead, we suspect that with some types of higher-level thinking skill tasks, dichotomous scoring may simply not be suitable (e.g., performance tasks, presentations, lab reports). Our work shows that either dichotomous or partial credit scoring methods are acceptable when assessing constructed response tasks requiring application of mathematical problem-solving skills where there is overlap and justification to use either scoring method (Brookhart & Nitko, 2008). With this in mind, we recommend additional research in this area with other mathematical constructed response items that could effectively be scored

using either method to see if findings are replicable based on the specific mathematical domain and task. We also recommend that future research test the potential differential impact of the two scoring methods for other relevant subgroups identified in *The Standards for Educational and Psychological Testing* (AERA et al., 2014) that were not examined in our study (e.g., race and/or ethnicity, linguistic background, disability status).

ORCID

Toni A. May https://orcid.org/0000-0001-7264-5607
Kristin L. K. Koskey https://orcid.org/0000-0002-9473-7956
Jonathan D. Bostic https://orcid.org/0000-0003-2506-0491
Lance M. Kruse https://orcid.org/0000-0003-1706-2286
Gabriel Matney https://orcid.org/0000-0001-8361-518X

REFERENCES

Albanese, M., & Sabers, D. (1998). Multiple true-false items: A study of inter-item correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement*, *25*, 111–123.

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

Bashkov, B., Mattison, K., & Hochstein, L. (2021, March). *IXL design principles: Core features groupnded in learning science research* [White paper]. https://www.ixl.com/research/IXL_Design_Principles.pdf

Bauer, D., Holzer, M., Kopp, V., & Fischer, M. R. (2011). Pick-N multiple choice-exams: A comparison of scoring algorithms. *Advances in Health Sciences Education*, *16*(2), 211–221. https://doi.org/10.1007/s10459-010-9256-1

Bendig, A. W. (1954). Reliability and the number of rating-scale categories. *Journal of Applied Psychology*, *38*(1), 38–40. https://doi.org/10.1037/h0055647

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Record*, *110*(3), 608–645.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates.

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., Pape, S., & Jacobbe, T. (2011). Validating two problem-solving instruments for use with sixth-grade students. In L. Wiest & T. Lamberg (Eds.), *Proceedings of the 33rd annual meeting of the north American chapter of the International Group for the Psychology of mathematics education* (pp. 756–763). Reno.

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. D., Sondergeld, T. A., Matney, G., Stone, G., & Koskey, K. L. (2022). Three steps forward: Validity evidence for the PSM. In D. Olanoff, K. Johnson, & S. Spitzer (Eds.), *Proceedings of the 43rd annual meeting of the north American chapter of the International Group for the Psychology of mathematics education* (pp. 26–30). Philadelphia.

Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Pearson.

Brookhart, S. M., & Nitko, A. J. (2019). *Assessment and grading in classrooms* (8th ed.). Pearson/Merrill Prentice Hall.

Charles, R. I., & Lester, F. K. (1984). An evaluation of a process-oriented instructional program in mathematical problem solving in grades 5 and 7. *Journal for Research in Mathematics Education*, *15*(1), 15. https://doi.org/10.2307/748985

Common Core State Standards Initiative. (2010). *Common core standards for mathematics*. National Governors Association Center for Best Practices & Council of Chief State School Officers.

Duncan, P. W., Bode, R. K., Min 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*(7), 950–963.

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*(1), 79–96. https://doi.org/10.1207/s15324818ame0201_5

Goldin, G. (2002). Representation in mathematical learning and problem solving. In L. D. English & M. G. Bartolini Bussi (Eds.), *Handbook of international research in mathematics education* (2nd ed., p. 925). Routledge.

Grizzle, R. (2011). Wechsler intelligence scale for children. In S. Goldstein & J. A. Naglieri (Eds.), *Encyclopedia of child behavior and development* (4th ed.). Springer.

Grunert, M. L., Raker, J. R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: Development of a rubric for rating partial credit. *Journal of Chemical Education*, *90*(10), 1310–1315. https://doi.org/10.1021/ed400247d

Jiao, H., Liu, J., Haynie, K., Woo, A., & Gorham, J. (2012). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement*, *72*(3), 493–509. https://doi.org/10.1177/0013164411422903

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130–144.

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

Lau, A., & Wang, T. (1998). *Comparing and combining dichotomous polytomous items with SPRT procedure in computerized classification testing*. Presented at the Annual Meeting of the American Educational Research Association.

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

Linacre, J. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions*, *12*(2), 636.

Linacre, J. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Linacre, J. (2012). *Winsteps (Version 3.74)*. Winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

McMillan, J. H. (2011). *Classroom assessment: Principles and practice for effective standards-based instruction* (5th ed.). Pearson.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research, and Evaluation*, *7*, 25. https://doi.org/10.7275/gcy8-0w24

Mertler, C. A. (2003). *Classroom assessment: A practical guide for educators*. Pyrczak Publishing.

Meyer, J. P., & Dahlin, M. (2022). *MAP growth theory of action [White paper]*. NWEA. https://www.nwea.org/content/uploads/2022/03/MAP-Growth-theory-of-action_NWEA_whitepaper.pdf

Mullis, I. V. S., Martin, M. O., & Loveless, T. (2016). *20 years of TIMSS: International trends in mathematics and science achievement, curriculum, and instruction*. https://timssandpirls.bc.edu/timss2015/international-results/timss2015/wp-content/uploads/2016/T15-20-years-of-TIMSS.pdf

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (2014). *Principles to action: Ensuring mathematical success for all*. National Council of Teachers of Mathematics.

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

Popham, W. J. (2014). *Classroom assessment: What teachers need to know* (7th ed.). Pearson Education.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Instiut.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests (expanded ed)*. University of Chicago Press.

Renaissance Learning. (2022). *STAR math*. https://www.renaissance.com/products/star-math/

Ripkey, D. R., Case, S. M., & Swanson, D. B. (1996). A "new" item format for assessing aspects of clinical competence. *Academic Medicine: Journal of the Association of American Medical Colleges*, *71*(10 Suppl), S34–S36.

Rogers, W. T., & Ndalichako, J. (2000). Number-right, item-response, and finite-state scoring: Robustness with respect to lack of equally classifiable options and item option independence. *Educational and Psychological Measurement*, *60*(1), 5–19. https://doi.org/10.1177/00131640021970330

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

Schrank, F. A., & Wendling, B. J. (2018). The woodcock-Johnson IV: Tests of cognitive abilities, tests of oral language, tests of achievement. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 383–451). The Guildford Press.

TIMSS. (2015). *TIMSS 2015 encyclopedia*. TIMSS & PIRLS International Study Center.

Verschaffel, L., De Corte, S., Lasure, S., van Vaerenbergh, G., Bogaertsm, 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., & de Corte, E. (2000). *Making sense of word problems*. Swets & Zeitlinger.

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

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.

How to cite this article: May, T. A., Koskey, K. L. K., Bostic, J. D., Stone, G. E., Kruse, L. M., & Matney, G. (2023). Examining how using dichotomous and partial credit scoring models influence sixth-grade mathematical problem-solving assessment outcomes. *School Science and Mathematics*, 1–14. https://doi.org/10.1111/ssm.12570