JOURNAL OF APPLIED MEASUREMENT, *23*(3/4), xx-xx

Validity and Test-Length Reduction Strategies for Complex Assessments

Lance M. Kruse
invontics, LLC.

Gregory E. Stone
University of Toledo

Toni A. May
Drexel University

Jonathan D. Bostic
Bowling Green State University

Lengthy standardized assessments decrease instructional time while increasing concerns about student cognitive fatigue. This study presents a methodological approach for item reduction within a complex assessment setting using the *Problem Solving Measure for Grade 6 (PSM6)*. Five item-reduction methods were utilized to reduce the number of items on the *PSM6*, and each shortened instrument was evaluated through validity evidence for test content, internal structure, and relationships to other variables. The two quantitative methods (Rasch model and point-biserial) resulted in the best psychometrically performing shortened assessments but were not representative of all content subdomains, while the three qualitative (content preservation) methods resulted in poor psychometrically performing assessments that retained all subdomains. Specifically, the ten-item Rasch and ten-item point-biserial shortened tests demonstrated the overall strongest validity evidence, but future research is needed to explore the psychometric performance of these versions in a new independent sample and the necessity for subdomain representation. Implications for the study provide a methodological framework for researchers to use and reduce the length of existing instruments while identifying how the various reduction strategies may sacrifice different information from the original instrument. Practitioners are encouraged to carefully examine to what extent their reduced instrument aligns with their pre-determined criteria.

Keywords: item reduction, validity evidence, psychometric, assessments, Rasch

Lance M. Kruse https:// orcid.org/0000-0003-1706-2286
Jonathan D. Bostic https:// orcid.org/0000-0003-2506-0491
Requests for reprints should be sent to Dr. Lance Kruse, 3413 Opequon Drive, Raleigh, NC 27610, USA; lancek@invontics.com

Introduction

Assessment is an area of critical importance within many educational systems around the world (Holloway, 2003; Organization for Economic Cooperation and Development, 2013). While vital elements in evaluation, standardized student assessments require a great deal of classroom time to administer, raising concerns amongst educators who often already feel time pressure to cover necessary content.

Two primary areas of concern regarding test administration are often cited when criticizing assessment programs. First, standardized test administration in the classroom reduces instructional time (Rentner et al., 2016). Noting a decided pressure to address content standards, many teachers dislike the frequency and duration of instructional time that standardized tests consume (Ferguson et al., 2017). Second, longer assessments also introduce the problem of cognitive fatigue by requiring students to engage in demanding tasks for an extended period, potentially stressing their mental resources (Sievertsen et al., 2016). Cognitive fatigue has been found to negatively affect standardized test performance in general, and germane to this study, mathematics and problem-solving outcomes (Gillmor et al., 2015; Nagane, 2004; Sievertsen et al., 2016). Fatigue also appears to reduce cognitive flexibility, which hinders students' ability to adapt to environmental changes when completing tasks (Plukaard et al., 2015). While the implications of cognitive fatigue remain unclear, it has been well-documented that test length increases students' self-reported levels of fatigue (Ackerman & Kanfer, 2009).

The type of items on an assessment may contribute to perceived cognitive fatigue (Kinsman & Weiser, 1976; Yeh & Wickens, 1988). High-level complex items, which require the use of reasoning to identify novel solution strategies connected with conceptual understanding, are more cognitively demanding than less-complex routine tasks, such as those on typical standardized multiple-choice exams (Kilpatrick et al., 2001; National Council Teachers of Mathematics [NCTM], 2014). Unfortunately, most studies exploring test length and cognitive fatigue involved monotonous, repetitive tasks (e.g., Ackerman & Kanfer, 2009; Davis, 1946; Jensen et al., 2013; Lee et al., 2015; Martyn, 1913), and because most were conducted using undergraduate students, it is not clear to what degree findings from prior research apply to K-12 students.

One proposed solution for time-intensive assessments is administration across consecutive days but in smaller time increments (e.g., five days of testing for 30 minutes each rather than a single day of testing). The limited research conducted in this area is unclear. Pope and Fillmore (2015) noted an increase in performance but only when time intervals between testing were at least ten days, which is administratively problematic in a traditional classroom. If time plays a critical role in student performance on assessments, it may make more sense to employ shorter, well-targeted tests that do not consume undue amounts of instructional time. That said, shortening already-developed instruments is not a straightforward task.

Strategies for item reduction are well-established in some professional disciplines (e.g., Computer Adaptive Testing [CAT]; Stafford et al., 2019), medical diagnostic assessments (Bilker et al., 2014), psychological clinical diagnosis (Fokkema et al., 2014), perceptual and attitudinal surveys (Ward et al., 2018). However, strategies for shortening classroom paper and pencil, non-adaptive tests, have not been adequately explored through rigorous methodological approaches (Goetz et al., 2013; Weiland et al., 2012). Goetz and colleagues (2013) argued that developers should: (1) document the validity of the results and interpretation of the original assessment and the purpose of shortening it, (2) consider the conceptual model of the assessment, (3) preserve content validity, (4) preserve psychometric properties, (5) justify the selection of each item, and (6) validate the short-form assessment with an independent sample. Similarly, Bostic and Sondergeld (2015), building on *The Standards*

Table 1

Framework Connecting Five Sources of Validity Evidence and Associated Evaluation Method

Source of validity evidence	Example evaluation methods
Test content	Qualitative expert panel review
Internal structure	Rasch psychometric analysis (reliability, separation, targeting)
Response processes	Cognitive interviews with test takers*
Relations to other variables	Total score correlations Differences based on gender (t-test) Differences based on ability level (ANOVA) Quartile placement
Consequences of testing	Cognitive interviews with test takers*

Note. These sources of validity evidence were not replicated as they were conducted with the same items during the original validation study and would likely not produce any differential findings (see Bostic & Sondergeld, 2015).

for Educational and Psychological Testing (American Educational Research Association [AERA] et al., 2014), described the five major sources of validity required in educational assessment (Table 1) and offered one framework to gather validity evidence. The current study builds upon that framework and explores both content (qualitative) and psychometric (quantitative) approaches to item reduction.

Content (Qualitative) Approaches

Content strategies for shortening instruments place the focus on retaining the original content blueprint for the shortened instrument (i.e., preserving content validity; Goetz et al., 2013). While the argument for retaining more or less identical content on the shortened instrument has been argued for many years (see Coste et al., 1997), models for best preserving test content during the item-reduction process are not well defined. Expert panels are often utilized in determining content validity during instrument construction (AERA et al., 2014). Yet, it is unclear how to best structure similar panels for assisting in item reduction while maintaining content validity (Goetz et al., 2013). Furthermore, the question becomes what content is being preserved? If the holistic construct is equivalent (at the macro level), is preservation of specific tasks (at the micro level or sub-domains) really required to maintain content validity?

Psychometric (Quantitative) Approaches

Psychometric strategies for item reduction focus on identifying best-performing items for retention on shortened instruments without *solely* focusing on the content blueprint of the original assessment (Goetz et al., 2013). Popular approaches to quantitative item reduction include the use of confirmatory factor analysis (Wolverton et al., 2018), item-total (point-biserial) correlations (Beaton et al., 2005), and Rasch (1960, 1980) models (Erhart et al., 2010; Nijsten et al., 2006). Although there is no consistent best practice model for psychometrically-based item reduction, these three strategies have shown promise. A confirmatory factor analysis strategy identifies and includes the highest loading items on the latent trait being measured. These values are useful when a factor structure is known and when sample size is adequate (Goetz et al., 2013). An item-total (point-biserial) correlation strategy involves the selection of items with the highest correlation to the total score (Beaton et al., 2005). While different than the confirmatory factor analysis, the use of item-total correlations may be problematic because such statistics are sample dependent and may not replicate for other samples. The use of Rasch models operationalizes the shortened assessment through its traditional explication of the construct via items arranged along a variable continuum. While variations across item-reduction strategies using the Rasch model exist, most suggest item removal based on variable construction and item performance indicators including fit and error (Beaton et al., 2005; Erhart et al., 2010; Nijsten et al., 2006).

Complex Assessments

Research relative to item reduction strategies has focused on objective item types, including dichotomously scored multiple-choice questions and evaluative items utilizing Likert-type rating scales. Complex assessments require students to engage in reasoning and problem solving to solve a problem involving many elements (NCTM, 2014). They also require significantly more time for students to complete, and as such, fewer are included in an instrument (Blum, 2015). Assessments with few items may pose significant psychometric challenges to the process of item reduction because most statistical inferences work best with greater numbers of items (Costello & Osborne, 2005). It is easier to balance content when more items are available from which to select the shortened instrument.

The Purpose of the Present Study

The current study compares and contrasts content and psychometric-based item-reduction strategies holistically, using a complex mathematics assessment to best understand both the practical classroom applications and theoretical imperatives relative to validity. A related secondary purpose is to determine if some subdomains of the construct may be removed without diminishing the instrument's capability to assess students' problem-solving abilities. One overarching research question was investigated: Are systematic best practices for reducing complex test length identifiable and implementable such that the outcomes from both longer and shorter assessments meet the five sources of validity evidence?

Methods

Instrumentation

Past data collected from the *Problem Solving Measure for Grade 6* (*PSM6*; Bostic & Sondergeld, 2015) were used in this study. The *PSM6* is a type of complex problem-solving assessment for middle-grade mathematics students aligned with the U.S. Common Core State Standards for Mathematics (CCSSM; Common Core State Standards Initiative [CCSSI], 2010). In response to the explication of mathematical problem solving as the first Standard for Mathematical Practice (SMP) in the CCSSM and highlighted in the Standards for Mathematics Content across the grade levels, Bostic and colleagues developed a set of *Problem Solving Measures* (*PSMs*) for middle school students aligned with the CCSSM (Bostic & Sondergeld, 2015; Bostic et al., 2017).

Although the *PSM6* is aligned with the five mathematical domains identified by the CCSSM, the *PSM6* is designed to measure students' problem-solving ability situated within mathematics more broadly that spans across the mathematical domains as described in the first SMP. Problem solving is not necessarily domain-specific but requires students to develop novel solution strategies, self-evaluate their strategy, utilize various mathematical representations, and verify their solutions across various mathematical contexts (CCSSI, 2010). Unfortunately, most mathematical tasks do not promote such reasoning and problem-solving skills and are commonly referred to as routine tasks (Kilpatrick et al., 2001). Instead, tasks that promote problem solving "encourage reasoning and access to the mathematics through multiple entry points, including the use of different representations and tools, and they foster the solving of problems through varied solution strategies" (NCTM, 2014, p. 17). Such tasks require students to draw on their prior knowledge and experience to assist in solving the problem (NCTM, 2014), yet extend beyond simple comprehension. Tasks that require the use of prior (non-academic) knowledge

may encourage students to implement novel problem-solving approaches (Matney et al., 2013; Palm, 2008), thus supporting the need for realism in problem solving. *Principles and Standards* (NCTM, 2000) provides a similar characteristic of problems by describing them as tasks without a known solution method. Mathematical problems may then be described as open tasks (containing multiple solution pathways), realistic (based on student's prior experiences), and complex (requiring mathematical reasoning and problem solving to find an unknown solution method; Verschaffel et al., 1999). It is, therefore, necessary to pose such problems to students that allow for such a complex and rigorous process.

Building upon the work by Verschaffel and colleagues (1999), the *PSMs* pose cognitively demanding word problems that require students to use novel solution strategies situated in the real world to arrive at one of the various potential solutions. Such items resemble the non-routine tasks that encourage students to engage in problem solving, as defined by the National Council Teachers of Mathematics (NCTM, 2000, 2014) and the CCSSM (2010). The items on the *PSM6* are scored dichotomously (correct or incorrect). However, the possibility to provide partial credit for incorrect answers that utilize a correct mathematical strategy (representation or procedure) was recognized, and an empirical evaluation of scoring models on the *PSM6* has been performed (see May et al., 2023). Through the evaluation of internal structure and consequential validity, it was found that there were nearly identical psychometric properties and student problem-solving ability classification between both scoring models (May et al., 2023). The minimal psychometric benefits of the partial-credit model did not justify the extended time required to provide partial credit for incorrect answers that utilize a correct mathematical strategy. As a result, the dichotomous scoring model data were utilized in this study to align with the original design of the instrument and allow for direct comparisons with the validity evidence provided in the instrument's original validation study (Bostic & Sondergeld, 2015).

The *PSM6* was developed and utilized as one of the measures to evaluate a professional development grant that focused on mathematical problem solving for middle-grade mathematics teachers in one state in the Midwest. As such, the instrument is anticipated to be used by middle grades mathematics teachers as means to evaluate students' mathematical problem-solving ability as described by NCTM. The instrument should not be used for high-stakes decision-making within the classroom or school for either student or teacher. Each of the mathematical domains from the CCSSM is represented on each *PSM* to ensure content alignment. The *PSM6* includes 15 items administered across two class sessions for an average time of 75 minutes (thus allowing about five minutes per item). Figure 1 presents an example of an item taken from the *PSM6*. Validity evidence related to the *PSM6* (including psychometric evidence from the Rasch model) and arguments for score interpretation and use are provided in Bostic and Sondergeld (2015).

Figure 1

Sample PSM6 Item

A group of 150 tourists were waiting for a shuttle to take them from a parking lot to a theme park's entrance. The only way they could reach the park's entrance was by taking this shuttle. The shuttle can carry 18 tourists at a time. After one hour, everyone in the group of 150 tourists reached the theme park's entrance. What is the fewest number of times that the shuttle picked tourists up from the parking lot?

Validity and Test-Length Reduction Strategies

Research Design

Multiple shortened assessments were constructed based on each of the useful content and psychometric strategies presented, except for the factor analytic approach. Only five of the original 15 items on the *PSM6* satisfied the initial criteria for conducting a confirmatory factor analysis (i.e., tetrachoric correlations > 0.30 and communalities > 0.20). To utilize a factor analysis approach, a greater number of items must meet targeting requirements. Because items requiring complex thinking also require significant time for student completion, it is unclear whether any assessment heavily concentrated on complex problems could meet this goal, and, as a result, the factor analysis strategy was not able to be implemented. The other strategies (i.e., item-total correlations, Rasch, content only, content with item-difficulty performance information, and content with both item-difficulty and point-biserial performance information) were used, and their ability to produce identical or improved student diagnostic measures while meeting the five validity sources (AERA et al., 2014) on a shortened test were evaluated.

Using each of the five item-reduction strategies, the researchers attempted to create 5-item and 10-item assessments. The five mathematical sub-domains included on the *PSM6* helped determine the length of the shortened tests because they allowed for an even spread of item content using the content (qualitative) strategies. After construction, each shortened version was evaluated according to the five sources of validity evidence. An expert panel was used to evaluate the content validity of shortened instruments. Internal structure was explored through the use of several Rasch (1960, 1980) analyses indicative of student measure quality and consistency (i.e., item/person reliability/separation and item/person infit/outfit mean squares), which were compared across the original version of the *PSM6* and each shortened version. First, the baseline *PSM6* performance was compared to the validation study (Bostic & Sondergeld, 2015) to ensure the baseline performance was similar to the performance of the original instrument and that the data from the present study satisfied established psychometric thresholds. The baseline psychometric performance of the *PSM6* was obtained through a dichotomous Rasch model analysis using Winsteps version 4.0.1 (Linacre, 2017). Specifically, item and person separation values of 1.50 are considered acceptable, values of 2.00 are considered good, and values of 3.00 are considered excellent (Wright & Masters, 1982), item and person infit and outfit mean-square values should range between 0.6 and 1.4 logits (Wright & Linacre, 1994), and persons and items with negative point-measure correlations may be removed from the analysis in order to create a more parsimonious measure of the latent trait under investigation (Boone et al., 2014). Since Winsteps was used for the Rasch analysis, it is important to note that Winsteps provides measures of reliability and separation for both persons and items, which may differ from other software.

Separation is the ability of the instrument to "represent a person's ability as independent of the specific test items and item difficulty as independent of specific samples within standard error estimates" (Bond & Fox, 2015, p. 349). The separation indicates the "spread" of items and persons such that the larger than spread, the easier it is to meaningfully distinguish items and persons from each other (Wright & Stone, 2004, p. 49). Person separation can be used to classify people, while item separation is used to verify the item hierarchy (Linacre, 2022). Reliability (i.e., separation reliability) is a correlation coefficient that models the "reproducibility of relative measure location" (Linacre, 2022, para. 5). In other words, the greater the reliability, the more likely it is that persons and items with reported higher measurement scores do have higher measures than those with lower measurement scores. From these perspectives, the person and item reliability and separation indices can be used to evaluate the performance of the instrument.

Relationships to other variables were

explored through total score bivariate Pearson correlations between the original and the shortened assessments, comparison of quartile student performance groups, and exploring differences based on gender and ability level. The quartile comparison involved the creation of two quartile groups such that the first quartile grouping was based on their performance on the original *PSM6*, and the second quartile grouping was based on their performance on each of the shortened *PSM6s*. After adding and subtracting two times the corresponding standard error of measurement (SEM) values to the student ability measures, the quartile placement of students was compared between the two instruments. Although it was expected to observe differences in the absolute classification of student quartile placement, only students who were in a quartile beyond two SEMs were identified as being in a statistically different quartile. The percentage of students who remained in the same or statistically similar quartile (after adding/subtracting two times the SEM) for both instruments was determined and reported. A higher percentage of students who remain in the same or statistically similar quartile is indicative of more decision consistency between the shortened version and the original *PSM6* as compared to test versions with lower percentages of students who remain in the same or similar quartile. Differences based on gender identity were evaluated through a series of independent samples *t*-tests and differences based on teacher-perceived ability level were evaluated using a one-way ANOVA. Evidence from the response process and consequences of testing were previously explored through cognitive interviews with test takers; thus, were not replicated in the present study as the items themselves did not change.

Successful evidence of validity was suggested to occur when three goals were met. First, test content was sufficiently addressed if each test form appropriately represented the holistic domain. Second, internal structure was satisfied if the assessments demonstrated psychometric performance that was deemed acceptable across its various measures. Third, relationships to other variables were determined to be acceptable through the analysis of total-score correlations, test-taker quartile placements, no statistical differences based on gender identity, and significant differences based on teacher-perceived student ability level.

Sample and Data Collection

A student sample of 517 sixth-grade students across eight purposefully selected Midwest schools and 16 classrooms was used for the quantitative analyses performed with the instrument. Purposeful selection of participants was intended to highlight socioeconomic and geographic diversity (i.e., urban, suburban, and rural). Students were tested during the last month of the school year to ensure all content areas had been addressed. After administration, teachers were asked to report the student's gender identity (male, female) and general academic ability level (below average, average, above average) by considering the student's other mathematical assessments, state test data, and a holistic assessment of the student's mathematics work. A relatively equal distribution of males and females was found, and there were more students classified as average than above or below average (see Table 2).

Table 2

Student Demographics (N = 517)

Student demographic variables *Values of variable*	Frequency (Percentage)
Gender	
Female	243 (47.0%)
Male	261 (50.5%)
Not provided	13 (2.5%)
Teacher reported student mathematics ability	
Above average	78 (15.1%)
Average	157 (30.4%)
Below average	115 (22.2%)
Not provided	167 (32.3%)

For content (qualitative) item reduction, an expert panel was convened through purposive sampling. The expert panel consisted of two tenured mathematics education professors and four middle-grade mathematics teachers. Inclusion criteria were that all teachers were sixth-grade mathematics teachers with teaching licenses from a four-year accredited institution and had at least three years of teaching experience. Exclusion criteria used were intervention specialists or other classroom aides who do not design and provide the student's mathematics lessons. Most panelists reported being very or extremely knowledgeable about the SMPs and the five domains in the CCSSM. Two panelists reported being somewhat knowledgeable about both the SMPs and domains in the CCSSM.

Shortened Test Development Process

As expressed, two psychometric (point-biserial and Rasch) and three content-based strategies (content only, content with item-difficulty performance information, and content with both item-difficulty and point-biserial performance information) were employed to create the shortened test forms. Each strategy is described in the following sections.

Shortened test forms using the point-biserial test reduction strategy were constructed by retaining items with the highest point-biserial values. To evaluate the threat of multicollinearity, point-biserial values were calculated between each item and the total person measure score from the baseline Rasch model analysis. Also, each item was added to a linear regression model as independent variables with the person measure score as the dependent variable to derive the Variance Inflation Factor (VIF). After evaluating the threat of multicollinearity, the ten-item shortened test form (*PB10*) was created by retaining the ten items with the highest point-biserial values. To create the five-item shortened test form (*PB5*), the five items with the highest point-biserial values were retained.

Construction of the shortened test forms using the Rasch-item reduction methodology was first guided by evaluating item infit and outfit mean-squares (0.6 to 1.40 logits), Winsteps-generated item point-measure correlation (must be positive), and item difficulty. Since the *PSM6* was originally developed and evaluated through the Rasch model, an emphasis was placed on the appropriate targeting of item difficulty to person ability (Bond & Fox, 2015). Each item's difficulty value was first compared to the range of person ability values. Items with difficulty values that fell between the person mean plus and minus two times the person standard deviation were considered well-targeted for persons. The item difficulty indices were not fixed from the previous calibration, and thus item difficulty indices were recalibrated for each analysis. After appropriately targeted items were retained that demonstrated acceptable fit, items were evaluated based on their impact on the overall item and person reliability and separation indices. Through an iterative process, the item which least impacted person and item reliability and separation values was removed, and the remaining items re-analyzed through Rasch. Each successive iteration allowed for the removal of the single weakest item, and the process continued until only five items remained, producing the five-item shortened test form (*R5*).

Each content item-reduction strategy was guided first by calculating a Content-Validity Ratio (CVR; Lawshe, 1975). An item's CVR is calculated using the following formula: $(N_e - N/2) / (N/2)$, where N_e is the number of panelists indicating "essential" and N is the total number of panelists. Because only six panelists completed the review, the CVR was not compared to a critical value, which would have required unanimous selection across items. Instead, items were ranked according to CVR within each domain. Panelists were required to conduct three different rating processes by utilizing different item metrics to inform their decisions for each of the content-based strategies. The content-only strategy required panelists to rate whether an item was essential by only considering the item and its associated content domain. To combine aspects of the quantitative and qualitative item-reduction processes, the content with item difficulty

strategy required panelists to rate the items when considering the item difficulty statistics in addition to content. Finally, to develop shortened assessments using the content with both item difficult and point-biserial strategy, panelists were provided with point-biserial values in addition to content and item difficulty. For each domain, the one or two items with the highest CVRs were identified for retention for each shortened test version.

Results

To compare the performance of subsequently shortened assessments, baseline performance of the *PSM6* was first established. The evaluation of person performance focused on person fit and person point-biserial values, which are provided in the person fit output table by Winsteps (Linacre, 2012). A total of 49 persons were initially removed (47 misfitting persons with outfit MNSQ values greater than 2.0 and 2 persons with negative point-biserial correlations). In addition, perhaps because of the relative difficulty of the assessment, 127 students demonstrated minimum measure (extreme) scores, did not contribute to the construct definition and were therefore removed and the analysis re-run. A final sample of 341 students was used in all subsequent analyses, including the item-reduction process and evaluation of validity evidence for each shortened instrument.

Items on the *PSM6* were analyzed by investigating their fit, difficulty, and targeting. All 15 items demonstrated appropriate fit (infit and outfit mean squares between 0.60 and 1.40 logits) and positive point-biserial values (Bond & Fox, 2015). Therefore, all items were deemed to be functioning appropriately. Item ordering on the Wright map was evaluated and compared to the item ordering presented in the validation study (Bostic & Sondergeld, 2015). The consistency of item ordering between this baseline *PSM6* analysis and the validation study suggests that the construct was operationalized effectively and in accordance with information presented in the validation study (see Appendix). Items were generally more difficult than the average student's level of mastery, which was expected given the difficulty of the construct and was similar to the validation study (Bostic & Sondergeld, 2015). The *PSM6* displayed excellent item reliability (0.95) and item separation (4.29) and acceptable person reliability (0.70) and person separation (1.50). After the evaluation of the baseline instrument, the shortened tests were created.

Shortened Tests Results

The point-biserial item-reduction process first required an evaluation of multicollinearity. Since each item produced point-biserial values greater than 0.80 and each item's VIF was close to 1 (ranging from 1.072 to 1.45), the threat of multicollinearity was minimized. The ten-item and five-item tests were identified by retaining the items with the ten and five highest point-biserial values, respectively. Items retained by each strategy are presented in Table 3.

As expected, all 15 items on the *PSM6* demonstrated acceptable fit and point-measure correlations, which required the Rasch-item reduction process to focus on item difficulty targeting. Ten items were identified within the range of the person mean plus and minus two times the person standard deviation. Because the remaining number of items was ten, the ten-item shortened test form (*R10*) was identified. Since the resulting ten items demonstrated appropriate fit, point-measure correlations, and item difficulty targeting, the iterative removal of items from the set of ten items was performed. Five items were retained that had the least impact on item and person reliability and separation values to create the five-item shortened test from (*R5*).

The construction of the content tests utilized the CVR values for each item (see Table 4). Each domain had at least one item with a higher CVR than the other items within that domain. However, panelists tended to agree on the remaining items, and it was not possible to identify a second-highest item to create a 10-item test. The exception to this occurred

VALIDITY AND TEST-LENGTH REDUCTION STRATEGIES

Table 3

Item Specification per Test Form

Question #	domain	Test form						
		PB10	*PB5*	*R10*	*R5*	*C6*	*CD6*	*CDPB5*
1	Statistics probability	X	-	X	X	-	-	-
2	Geometry	-	-	-	-	X	X	X
3	Number sense	-	-	X	-	X	X	X
4	Ratio and proportion	X	X	X	X	X	X	X
5	Number sense	X	X	X	X	-	-	-
6	Expressions and equations	X	-	X	-	X	X	X
7	Geometry	-	-	-	-	-	-	-
8	Statistics probability	X	-	X	-	-	-	-
9	Expressions and equations	X	X	X	-	-	X	-
10	Ratio and proportion	X	-	-	-	-	-	-
11	Number sense	X	X	X	-	-	-	-
12	Ratio and proportion	X	-	X	X	-	-	-
13	Statistics probability	-	-	-	-	X	X	X
14	Geometry	-	-	-	-	-	-	-
15	Ratio and proportion	X	X	X	X	X	-	-

Note. An "X" indicates an item was retained for that test form, whereas a "-" indicates the item was omitted. Test forms are abbreviated in the following manner: PB10 = Point-Biserial 10-Item Test, PB5 = Point-Biserial 5-Item Test, R10 = Rasch 10-Item Test, R5 = Rasch 5-Item Test, C6 = Content 6-Item Test, CD6 = Content & Difficulty 6-Item Test, CDPB5 = Content, Difficulty, and Point-Biserial 5-Item Test.

Table 4

CVR Values for Each Item Derived From Each Qualitative Item-Reduction Methodology

Question #	Domain	Test form		
		C6	CD6	CDPB5
1	Statistics and probability	–0.67	–0.67	–0.67
8	Statistics and probability	–0.67	–0.67	–0.33
13	Statistics and probability	0.67	0.67	0.33
2	Geometry	1.00	0.67	0.67
7	Geometry	0.33	0.33	0.33
14	Geometry	0.33	0.33	0.00
3	Number sense	0.00	0.00	0.33
5	Number sense	–0.67	–0.33	0.00
11	Number sense	–0.67	–0.33	0.00
4	Ratio and proportion	0.33	0.33	0.33
10	Ratio and proportion	–0.33	–0.33	0.00
12	Ratio and proportion	–0.33	–0.33	–0.33
15	Ratio and proportion	0.33	0.00	0.00
6	Expressions and equations	0.33	0.00	0.67
9	Expressions and equations	0.00	0.00	0.00

Note. Test forms are abbreviated in the following manner: C6 = Content 6-Item Test, CD6 = Content & Difficulty 6-Item Test, CDPB5 = Content, Difficulty, and Point-Biserial 5-Item Test. Items are grouped by their domain for easier comparison of CVR values within each domain.

relative to two items in Ratios and Proportions (RP4 and RP15), which were tied for the highest CVR. The content-only strategy, which focuses exclusively on the CVR, resulted in a shortened test version where most subdomains were represented by one item, except for Ratios and Proportions, which was represented by two items. The approach resulted in a single, six-item shortened test (*C6*).

When panelists were presented with item difficulty statistics in addition to content (the content with item difficulty strategy) their ratings of item importance changed slightly. The inclusion of item difficulty eliminated the tie across the two Ratio and Proportions items from the content-only strategy but created a new tie within Expressions and Equations. Additionally, the panelists again universally agreed on the second-highest CVR, preventing the creation of a 10-item shortened test. As a result, only one item per domain was retained except in the category of Expressions and Equations, resulting in a six-item shortened test (*CD6*).

For the content with both item difficulty and point-biserial strategy, panelist ratings only marginally changed. There was no longer a tie within any domain, allowing for one item per domain. Again, because of the level of agreement across and within multiple domains, the construction of a 10-item test was not possible, however, the five-item shortened test was constructed (*CDPB5*).

Validity Evidence Based on Test Content

The most essential items per domain were identified to evaluate the validity evidence based on test content. From the third round of reviews by the expert panel, the five items with the highest CVR values per domain were identified as the most essential items per domain. Each test form was then evaluated by identifying the number of items included from each domain as well as how many of those included items were the essential item per domain (see Table 5).

With the exception of the ratios and proportions domain, the psychometrically-derived tests did not include items considered to be essential by the panelists across all of the five domains. For example, no geometry items were included in any of the quantitative test forms. Additionally, none of the Statistics and Probability items were deemed essential. As a result, the psychometrically-defined test forms do not represent the five domains completely. Conversely, each of the content (qualitative) shortened tests represented items from each of the five sub-domains, and therefore, from this perspective, demonstrated strong test content validity evidence.

Table 5

Items Included per Domain for Each Test Form

Test form	Items per domain: # included (# Essential)				
	Statistics & probability	Geometry	Number sense	Ratios & proportions	Expressions & equations
PB10	2 (0)	0 (0)	2 (0)	4 (1)	2 (1)
PB5	0 (0)	0 (0)	2 (0)	2 (1)	1 (0)
R10	2 (0)	0 (0)	3 (1)	3 (1)	2 (1)
R5	1 (0)	0 (0)	1 (0)	3 (1)	0 (0)
C6	1 (1)	1 (1)	1 (1)	2 (1)	1 (1)
CD6	1 (1)	1 (1)	1 (1)	1 (1)	2 (1)
CDPB5	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)

Note. Test forms are abbreviated in the following manner: PB10 = Point-Biserial 10-Item Test, PB5 = Point-Biserial 5-Item Test, R10 = Rasch 10-Item Test, R5 = Rasch 5-Item Test, C6 = Content 6-Item Test, CD6 = Content & Difficulty 6-Item Test, CDPB5 = Content, Difficulty, and Point-Biserial 5-Item Test.

VALIDITY AND TEST-LENGTH REDUCTION STRATEGIES

Validity Evidence Based on Internal Structure

The psychometric performance of each psychometrically and content-derived shortened test is presented in Tables 6 and 7 respectively, along with the *PSM6* baseline measures. All five quantitative assessments demonstrated excellent item reliability and separation but weaker person reliability and separation. Due to the fewer number of items on the assessments, low person reliability and separation values are expected as it is more difficult to meaningfully differentiate a person's ability with fewer items that span a limited range of difficulty (Linacre, 1993). The ten-item point-biserial and Rasch test forms performed best psychometrically, as the highest and second highest in terms of all indices, and only differed marginally from the baseline. Based on item performance, both of the ten-item psychometrically shortened test forms demonstrated acceptable validity evidence for internal structure. Although the five-item point-biserial and Rasch tests demonstrated high item reliability and separation, they also demonstrated very low person reliability and separation due to the few items included, which weakens the validity evidence for internal structure for these two instruments.

Content-based shortened tests were analyzed similarly. While each of the three test forms demonstrated excellent item reliability and separation, none of the content-based shortened assessments produced useful person measures. Person reliability and separation values were reported at 0, which was a result of not only the limited number of items included but particularly how poorly these included items were able to measure the sample. To explain these findings, separation is the ratio of true standard deviation to the square root of the average measurement error variance (Fisher, 1992), which can be expressed using G (Equation 1), where observed SD is the population standard deviation of the measure and RMSE is the square root of the average measurement error variance

$$G = \frac{\sqrt{(observed\ SD)^2 - RMSE}}{RMSE}. \qquad (1)$$

For each of the content tests, the observed SD was less than the RMSE, resulting in the square root of a negative value. For example, for the six-item content test, observed SD = 0.51 and RMSE = 1.40, which results in a negative value within the square root

$$\frac{\sqrt{0.51^2 - 1.40^2}}{1.40}. \qquad (2)$$

The square root of a negative value is an imaginary number and so the separation is

Table 6

Psychometric Performance of PSM6 (n = 341) for Baseline and Quantitative Test Forms

Rasch statistic/criteria	*Baseline*	*PB10*	*PB5*	*R10*	*R5*
Item reliability	0.95	0.99	0.97	0.98	0.99
Item separation	4.29	8.92	6.09	7.40	8.53
Person reliability	0.70	0.65	0.08	0.61	0.20
Person separation	1.51	1.38	0.29	1.25	0.50
Mean person score	–2.08	–1.07	–0.54	–0.90	–0.61
SD of person score	1.60	1.56	1.16	1.41	1.33
# Difficult items	4	3	0	0	0
# On-target items	11	7	5	10	5
# Easy items	0	0	0	0	0

Note. Test forms are abbreviated in the following manner: PB10 = Point-Biserial 10-Item Test, PB5 = Point-Biserial 5-Item Test, R10 = Rasch 10-Item Test, R5 = Rasch 5-Item Test.

Table 7

Psychometric Performance of PSM6 (n = 341) for Baseline and Qualitative Test Forms

Rasch statistic/criteria	*Baseline*	*C6*	*CD6*	*CDPB5*
Item reliability	0.95	0.91	0.91	0.94
Item separation	4.29	3.16	3.17	3.81
Person reliability	0.70	0.00	0.00	0.00
Person separation	1.51	0.00	0.00	0.00
Mean person score	–2.08	–0.29	–0.33	–0.21
SD of person score	1.60	0.93	0.92	0.73
# Difficulty items	4	2	2	2
# On-target items	11	4	4	3
# Easy items	0	0	0	0

Note. Test forms are abbreviated in the following manner: C6 = Content 6-Item Test, CD6 = Content & Difficulty 6-Item Test, CDPB5 = Content, Difficulty, and Point-Biserial 5-Item Test.

reported as 0. Since separation (G) is used in the calculation for reliability (Equation 3; Fisher, 1992), the content tests that have a separation of 0 will also have a reliability of 0

$$Separation\ reliability = \frac{G^2}{1+G^2}. \quad (3)$$

Practically, the intent of the separation index is to assess whether the precision of measurements separates people into class intervals that do not overlap. For these tests, the specific set of items retained for the content strategies was not able to reliably measure person ability across the continuum of the construct due to the few items included and the limited span of person ability (Linacre, 1993). The validity evidence for internal structure of the content-based tests was considerably weaker than the validity evidence for internal structure of the psychometrically-derived shortened tests, especially the ten-item psychometric tests.

Validity Evidence Based on Relationships to Other Variables

Differences based on gender identity were explored through a series of independent-samples t-tests for each test form with a Bonferroni-adjusted alpha. No significant differences in person measures based on gender identity were noted. Differences based on teacher-perceived ability level (above average, average, below average) were evaluated through a one-way ANOVA using person measure scores. All of the assessments significantly differentiated based on ability level ($p < 0.001$), with large effect sizes ($\eta^2 = 0.20$ to 0.32) indicating between 20 to 32% of the variance in person measure scores is accounted for by the students' teacher-perceived ability level. The large measure of effect size suggests that there is a strong relationship between student performance and teacher-perceived ability level and supports the idea that student performance on the instrument aligns with their expected ability level, providing additional validity evidence. Fisher's Least Significant Difference post hoc analysis identified each of the pairwise comparisons was significantly different except for the five-item qualitative test, which did not yield significant differences between average and below-average students. Such findings indicate that all forms of the shortened test, except for some cases in the five-item versions (between average and below-average students), meaningfully differentiated between students of different ability (i.e., above-average students > average students > below-average students).

Further convergent validity evidence was evaluated through bivariate Pearson Correlations between the person measure scores of each test form with the baseline. All four of the quantitative tests (i.e., Rasch five-item, Rasch ten-item, point-biserial five-item, point-biserial-ten item) demonstrated very strong, significant correlations between their test forms and the baseline ($r > 0.80$) with large effect sizes ($r^2 = 0.84$ to 0.98), and the three qualitative test forms (i.e., six-item content, six-item content and difficulty, and six-item content, difficulty, and point-biserial) demonstrated strong correlations with the baseline ($r = 0.723$ to 0.786). The large measure of effect size suggests that the scores from those four quantitative tests were strongly related to the baseline, supporting the idea that student performance between the instruments is very similar. All the shortened test forms resulted in person measure scores that were significantly, positively related to the original *PSM6*.

Lastly, to determine the consistency of outcome decisions, student quartile placement was compared between each test form and the baseline. As observed in Table 8, most test forms differed greatly in the absolute classification of students apart from the ten-item Rasch test, which resulted in an almost identical distribution of students. However, after accounting for the SEM, both the ten-item point-biserial and Rasch test forms resulted in 100% decision consistency with the original *PSM6* as all students were in the same or statistically similar quartile (students within two SEMs are considered to be in the similar quartile placement). Similarly, the five-item Rasch tests displayed reasonable decision consistency (96%). Conversely, the five-item point-biserial (90%), and all content-derived shortened tests were much less consistent (83% to 93%).

Table 8

Frequency and Percentage of Student Quartile Placement and Movement Across Each Test Form

Test version	Q1	Q2	Q3	Q4	# of students moved inside SEM	# of students moved outside SEM	% of students in the same or statistically similar quartile
Baseline	80	56	95	110	-	-	-
PB10	10	137	101	93	98	0	100%
PB5	100	0	150	91	103	33	90%
R10	80	56	96	109	1	0	100%
R5	49	111	83	98	107	14	96%
C6	84	0	111	146	128	35	90%
CD6	86	0	115	140	124	24	93%
CDPB5	107	0	127	107	107	58	83%

Note. Test forms are abbreviated in the following manner: PB10 = Point-Biserial 10-Item Test, PB5 = Point-Biserial 5-Item Test, R10 = Rasch 10-Item Test, R5 = Rasch 5-Item Test, C6 = Content 6-Item Test, CD6 = Content & Difficulty 6-Item Test, CDPB5 = Content, Difficulty, and Point-Biserial 5-Item Test.

Summary of Validity Evidence

When considering each source of validity evidence, each test form can be holistically evaluated by considering how well it addressed each type of validity evidence. No one shortened test demonstrated sufficient validity evidence for each type. That said, the ten-item Rasch and ten-item point-biserial test forms both demonstrated sufficient internal structure and relationship to other variables, suggesting they were the two best-performing shortened tests. However, the ten-item point-biserial test contained three items that were significantly too difficult for the sample of students (i.e., item difficulties were two standard deviations above the mean person ability), whereas the ten-item Rasch test included only items that fell within the appropriate item-difficulty range (i.e., within two standard deviations of the mean person ability). The notable exception to the Rash and point-biserial ten-item tests include the absence of subdomain representation for their

shortened versions. Even without including all subdomains, the ten-item Rasch test resulted in no significant shifts in person measure quartile groupings and the point-biserial test resulted in 98 non-significant quartile grouping shifts. Considering both the psychometric properties of the instrument and the resulting person measures, the Rasch ten-item shortened test was considered to be the best performing test, while the point-biserial ten-item test was rated as second-best, although neither contained absolute subdomain representation.

Discussion

Item-reduction methodologies within education generally, and complex assessments, in particular, are not well-elaborated (Goetz et al., 2013). It is important to explore approaches because complex item types are commonly used for measuring skills such as problem solving (Weiland et al., 2012). Identifying potential best practices for shortening educational assessments that include complex item types will assist future test developers in ensuring potentially erroneous decisions are not made from shortened assessments. The present study evaluated five commonly used methods (content and psychometric) for reducing existing assessments. Each item-reduction method was evaluated based on its ability to demonstrate acceptable sources of validity evidence for test content, internal structure, and response processes. The *PSM6* was used as an example of a previously developed modern assessment utilizing complex item types with a peer-reviewed validation study. The use of the *PSM6*, although being only one assessment, serves as an illustrative case for other similarly designed complex assessments for which current item-reduction methodologies are not established (Goetz et al., 2013). Results will be discussed within a three-part framework including assessment decision-making, test content, and test-specific considerations.

Assessment Outcome Decision Making

Both the Rasch and point-biserial strategies resulted in similarly performing tests with excellent levels of relationships to other variables and internal structure validity evidence. The Rasch strategy resulted in a test that performed marginally better than the point-biserial strategy in its ability to not only identify each item's contribution to measuring the construct but also by identifying the difficulty of items in relation to students' ability. Such findings align with previous research supporting the use of Rasch over point-biserial item-reduction strategies (Erhart et al., 2010). Evaluating the targeting of items is crucial in ensuring appropriate person and item reliability and separation (Linacre, 1997). To obtain better person reliability, it is more important to have items that are appropriately targeted to the sample of student ability (and to have a broader range of person ability) than to just have an abstractly large sample (Linacre, 1993). Similar to previous item-reduction studies (Weiland et al., 2012), Rasch was able to guide the selection of items for retention based on difficulty targeting to ensure appropriate person reliability. Both strategies likewise resulted in decision consistency outcomes that were statistically similar to the original *PSM6*. Use of either strategy resulted in more consistent results than any of the content-based strategies. While all content-derived test forms demonstrated excellent levels of test content validity evidence, they also demonstrated unacceptable levels of validity evidence for internal structure and relationships to other variables. Although a full set of ten- and five-item shortened tests could not be generated, those that were generated were very clearly deficient.

Test Content

While it is commonly agreed that test content validity evidence is the most fundamental source of validity evidence (AERA et al., 2014; Goetz et al., 2013), using content as a primary focus of item-reduction may not be as important as anticipated. The three content-based strategies emphasized maintaining the original content blueprint down to the level of subdomains. In doing so, however, the shortened test forms resulted in an instrument yielding different interpretations of scores as compared to the original instrument. Existing research that argues for the superiority of content-based item-reduction methods actually eliminated several subdomains in order to improve the measurement of the construct on the shortened test form (Beaton et al., 2005). Therefore, maintaining absolute subdomain representativeness may not be a critical factor to consider. Instead, it may be more important to ensure the test holistically aligns with the general construct.

In this more holistic way, every shortened test (across both sets of strategies) was considered to be aligned. The original *PSM6* (Bostic & Sondergeld, 2015) was developed with the Rasch model and ensured a unidimensional construct of mathematical problem solving was developed. Mathematical problem solving is not inherently domain-specific but instead is theorized as a process students utilize that is situated within various mathematical contexts. Specific emphasis was placed on aligning each item with one or two content standards, across all five domains. However, it was never suggested that each unique mathematical domain represented a specific aspect of problem solving that *must* be represented as part of the construct. Instead, all of the items on the *PSM6* work together to measure students' mathematical problem-solving ability and the content was usefully diverse. As a result, the Rasch item reduction strategy preserved the original unidimensional construct, even by excluding one domain. In instances where subdomains present a representational challenge, it is important to keep the larger holistic construct representation in mind, realizing that if decisions are consistent, it is likely that the same construct is being addressed in both the long and shortened versions.

Test Specific Considerations

Two distinct characteristics of assessments employing complex items may influence the success of each of the strategies deployed in this study: the number of items on the original instrument and the complexity of the construct being measured by the test. Relative to initial test length, because assessments employing complex items tend to include fewer items, there exists an *a priori* limited number of item combinations possible. Should these strategies be used on longer assessments, it is possible that some may provide more useful information than offered in the present study. Therefore, the generalizability of the findings of this study are limited to other similar complex assessments that utilize non-traditional assessment items. Differential results may be observed when using these item-reduction strategies on routine assessments that deploy common multiple-choice items that do not required mathematical problem solving as defined by NCTM (2000).

As noted earlier, when seeking to develop a well-functioning assessment, the alignment of item difficulty with student ability is a primary concern (Linacre, 1997). As seen in this study, more than 25% of the original sample was removed from the analysis as they did not answer any item correctly. These findings were not entirely surprising given the difficulty of mathematical problem-solving, but it limits the generalizability of these implications to students who are able to engage in this type of problem solving. The use of the *PSM6* is recommended to be used in classrooms where mathematical problem-solving skills are being fostered by teachers, which may not be applicable to every classroom and every student. Item-reduction strategies used to reduce the length of assessments that include difficult items measuring difficult constructs face an uphill challenge. Highly demanding constructs measured with few items require significant attention be paid to the strengths and psychometric properties of each item.

Limitations and Future Research

While the present study employed a rigorous methodological design as has been

lacking in other test-reduction studies (Goetz et al., 2013), certain limitations remain. The first is that the various item-reduction methodologies were used on only one assessment that was designed for measuring complex cognitive processes. Some outcomes from the item-reduction methodologies would likely vary should a different test, especially one that does not measure a complex cognitive process, be used. It is recommended that other practitioners apply the item-reduction strategies discussed in this study with other instruments and determine if there is a similar pattern of findings as presented here. Also, while the authors have found that a partial-credit scoring model, rather than a dichotomous scoring model, for this instrument does not change the psychometric performance of the instrument (May et al., 2023), it is important to consider how different scoring methods may have differential impacts during the item-reduction process and may be an area of future research. Given the dichotomous scoring of the assessment, the use of tetrachoric correlations was required for evaluating the appropriateness of items to include in the factor analysis process. Utilizing assessment with different scoring models may result in more items that are able to be retained in the factor analysis item-reduction process and provide another item-reduction strategy. Another limitation was the relatively small expert panel. A larger expert panel may have resulted in different outcomes and greater variance between panelists. The fact that the shortened assessments were not piloted using an independent sample is another limitation. Since 176 of the original 517 students were removed from the analysis due to misfit or extreme minimum scores, there is a limitation to the target population of this assessment. This assessment may not be entirely useful for students who have not received targeted instruction around mathematical problem solving as described by NCTM (2000). Also, while the ten-item Rasch test maintained decision consistency, this study did not assess the extent to which the construct being measured without geometry items resembles the construct being measured in the original 15-item test. The ten-item Rasch test and the ten-item point-biserial test also had person separation indices below the recommended value of 1.50, which hinders its ability to distinguish between groups of students. Therefore, the shortened instruments (as well as the original *PSM6*) should not be used for high-stakes decision making, but rather, as a classroom assessment to inform teachers about their students' mathematical problem-solving ability. Lastly, while it is recognized that the final step of the item reduction process would require an independent sample (Goetz et al., 2013), the present study sought only to identify which test forms might be the most appropriate to move forward to conduct pilot testing.

Future research should extend this present study by piloting the ten-item Rasch and point-biserial tests and the five-item Rasch and point-biserial tests in a new independent sample. Collecting new independent data for the shortened assessments would allow for the performance on the shortened tests to be more robustly compared to the original, including an evaluation of the construct being measured in the ten-item Rasch test as compared to the original instrument. Specifically, the potential impact of cognitive fatigue on both the original and the shortened instruments should be explored to better understand the implications of cognitive fatigue for middle school students engaging in mathematical problem solving. Such research may provide insightful thresholds for instrument length to guide future item-reduction processes. Relatedly, future research could explore the implications of these item-reduction strategies with other assessments with a different sample of students to determine the generalizability of these findings. While this study recalibrated item-difficulty indices during each Rasch analysis to derive the shortened instruments, future research could also explore the impact of using fixed item parameters for subsequent analyses. Additionally, it may be useful for future mathematics education scholars to develop and test additional geometric problem-solving tasks for inclusion in future instruments. Future research could explore if the construct of problem solving within geometry is intrinsically unique as compared to the other domains. In addition, the consequences of testing need to be evaluated through interviews with teachers and students relative to the potential implications of the assessments. A specific area of inquiry could include how teachers perceive the utility of a shortened instrument that does not include the geometry subdomain. Understanding how teachers use the outcome of the test to inform instruction, assessment, or other classroom decisions is a vital element in reflective assessment.

Final Thoughts

Assessments remain an important piece of the international educational system (Holloway, 2003; Organization for Economic Cooperation and Development, 2013) and assessment developers must continue to employ innovative techniques to meet the needs of an ever-changing and ever-more complex world. As educators have increased our focus on curating students' complex cognitive processes, our need to assess student ability across these skills has also increased. The use of item-reduction strategies to create shorter but equally effective assessments may help address this need by making such assessments more accessible to more classroom teachers. This study demonstrated that the use of Rasch-reported statistical performance information provided unique and unparalleled insight into guiding the item-reduction process. Best-practice strategies preserved the holistic latent trait and allowed for decisions consistent with the original assessment to be made. Assessments are measurement tools that make use of content reflecting the desired construct. As such, psychometrics must be strongly considered when making changes to an examination with a robust validity argument.

Acknowledgments

Ideas in this manuscript stem from grant-funded research by the National Science Foundation (NSF 1720646 & 1720661). Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163–181. https://doi.org/10.1037/a0015719

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

Beaton, D. E., Wright, J. G., Katz, J. N., & The Upper Extremity Collaborative Group. (2005). Development of the QuickDASH: Comparison of three item-reduction approaches. *The Journal of Bone & Joint Surgery*, *87*(5), 1038–1046. https://doi.org/10.2106/JBJS.D.02060

Bilker, W. B., Wierzbicki, M. R., Brensinger, C. M., Gur, R. E., & Gur, R. C. (2014). Development of abbreviated eight-item form of the Penn verbal reasoning test. *Assessment*, *21*(6), 669–678. https://doi.org/10.1177/1073191114524270

Blum, W. (2015). Quality teaching of mathematical modelling: What do we know, what can we do? In S. J. Cho (Ed.), *The proceedings of the 12th international congress on mathematical education* (pp. 73–96). Springer International Publishing. https://doi.org/10.1007/978-3-319-12688-3_9

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (Third edition). Routledge.

Boone W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*.

Springer. https://doi.org/10.1007/978-94-007-6857-4

Bostic, J., & Sondergeld, T. A. (2015). Measuring sixth-grade students' problem solving: Validating an instrument addressing the mathematics common core. *School Science and Mathematics*, *115*(6), 281–291. https://doi.org/10.1111/ssm.12130

Bostic, J., Sondergeld, T. A., Folger, T., & Kruse, L. (2017). PSM7 and PSM8: Validating two problem-solving measures. *Journal of Applied Measurement*, *18*(2), 151–162.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

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

Coste, J., Guillemin, F., Pouchot, J., & Fermanian, J. (1997). Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology*, *50*(3), 247–252. https://doi.org/10.1016/S0895-4356(96)00363-0

Costello, A., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment*, *10*(7), 1–9.

Davis, D. R. (1946). The disorganization of behaviour in fatigue. *Journal of Neurology, Neurosurgery & Psychiatry*, *9*(1), 23–29. https://doi.org/10.1136/jnnp.9.1.23

Erhart, M., Hagquist, C., Auquier, P., Rajmil, L., Power, M., & Ravens-Sieberer, U., & the European KIDSCREEN Group. (2010). A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of a Quality of Life scale for children and adolescents: Comparing Rasch item-fit and Cronbach's alpha analysis. *Child: Care, Health and Development*, *36*(4), 473–484. https://doi.org/10.1111/j.1365-2214.2009.00998.x

Ferguson, M., Kober, N., & Rentner, D. (2017). *What do teachers and district leaders think about state standards and assessments?* Center on Education Policy.

Fisher, W. P. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions*, *6*(3), 238.

Fokkema, M., Smits, N., Kelderman, H., Carlier, I. V. E., & van Hemert, A. M. (2014). Combining decision trees and stochastic curtailment for assessment length reduction of test batteries used for classification. *Applied Psychological Measurement*, *38*(1), 3–17. https://doi.org/10.1177/0146621613494466

Gillmor, S., Poggio, J., & Embretson, S. (2015). Effects of reducing the cognitive load of mathematics test items on student performance. *Numeracy*, *8*(1). https://doi.org/10.5038/1936-4660.8.1.4

Goetz, C., Coste, J., Lemetayer, F., Rat, A.-C., Montel, S., Recchia, S., Debouverie, M., Pouchot, J., Spitz, E., & Guillemin, F. (2013). Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. *Journal of Clinical Epidemiology*, *66*(7), 710–718. https://doi.org/10.1016/j.jclinepi.2012.12.015

Holloway, J. (2003). Using data to improve student achievement. *Educational Leadership*, *60*(5), 74–76.

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PLoS ONE*, *8*(8), Article e70270. https://doi.org/10.1371/journal.pone.0070270

Kilpatrick, J., Swafford, J., & Findell, B. (Eds). (2001). *Adding it up: Helping children learn mathematics*. The National Academies Press. http://site.ebrary.com/id/10038695

Kinsman, R., & Weiser, P. (1976). Subjective symptomatology during work and fatigue. In E. Simonson & P. Weiser (Eds.), *Psychological aspects and physiological correlates of work and fatigue* (pp. 336–405). Charles C. Thomas.

Lawshe, C. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lee, E., Garg, N., Bygrave, C., Mahar, J., & Mishra, V. (2015). *Can university exams be shortened? An alternative to problematic traditional methodological approaches* [Paper presentation]. European Conference on Research Methodology for Business and Management Studies, Kidmore End, Reading, UK.

Linacre, J. M. (1993). Rasch-based generalizability theory: Reliability and precisions (S.E.) nomogram. *Rasch Measurement Transactions*, *7*(1), 283–284.

Linacre, J. M. (1997). KR-20 / Cronbach alpha or Rasch person reliability: Which tells the truth? *Rasch Measurement Transactions*, *11*(3), 580–581.

Linacre, J.M. (2017). *Winsteps* (Version 4.0.1) [Computer Software]. Winsteps. https://www.winsteps.com/

Linacre, J. M. (2022). *A user's guide to Winsteps: Rasch-model computer programs*.

Lord, F. (1970). Some test theory for tailored testing. In W. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139–183). Harper and Row.

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205

Martyn, G. W. (1913). A study of mental fatigue. *British Journal of Psychology*, 1904–1920, 5(4), 427–446. https://doi.org/10.1111/j.2044-8295.1913.tb00073.x

Matney, G., Jackson, J., & Bostic, J. (2013). Effects of minute contextual experience on realistic assessment of proportional reasoning. *Investigations in Mathematics Learning*, *6*(1), 41–68.

May, T., Koskey, K., Bostic, J., Stone, G., Kruse, L., & Matney, G. (2023). Evaluating the differential impact of dichotomous and partial credit scoring models on student problem-solving assessment outcomes. *School Science and Mathematics*, *123*(2), 54–67. https://doi.org/10.1111/ssm.12570

Nagane, M. (2004). Relationship of subjective chronic fatigue to academic performance. *Psychological Reports*, *95*(1), 48–52. https://doi.org/10.2466/pr0.95.1.48-52

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

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

Nijsten, T. E. C., Sampogna, F., Chren, M.-M., & Abeni, D. D. (2006). Testing and reducing Skindex-29 using Rasch analysis: Skindex-17. *Journal of Investigative Dermatology*, *126*(6), 1244–1250. https://doi.org/10.1038/sj.jid.5700212

Organization for Economic Cooperation and Development. (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. https://www.oecd.org/education/school/synergies-for-better-learning.htm

Palm, T. (2008). Impact of authenticity on sense making in word problem solving. *Educational Studies in Mathematics*, *67*(1), 37–58. https://doi.org/10.1007/s10649-007-9083-3

Plukaard, S., Huizinga, M., Krabbendam, L., & Jolles, J. (2015). Cognitive flexibility in healthy students is affected by fatigue: An experimental study. *Learning and Individual Differences*, *38*, 18–25. https://doi.org/10.1016/j.lindif.2015.01.003

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

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

Rentner, D., Kober, N., Frizzell, M., & Ferguson, M. (2016). *Listen to us: Teacher views and voices*. Center on Education Policy.

Sievertsen, H. H., Gino, F., & Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences*, *113*(10), 2621–2624. https://doi.org/10.1073/pnas.1516947113

Stafford, R. E., Runyon, C. R., Casabianca, J. M., & Dodd, B. G. (2019). Comparing computer adaptive testing stopping rules under the generalized partial-credit model. *Behavior Research Methods*, *51*, 1305–1320. https://doi.org/10.3758/s13428-018-1068-x

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*(3), 195–229. https://doi.org/10.1207/s15327833mtl0103_2

Ward, T., Arnold, K., Cunningham, M. C., & Liljequist, L. (2018). Three validation studies of the personality assessment inventory short form. *Journal of Clinical Psychology*, *74*(12), 2264–2275. https://doi.org/10.1002/jclp.22677

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of the short form of a prekindergarten and kindergarten mathematics measure. *Educational Psychology*, *32*(3), 311–333. https://doi.org/10.1080/01443410.2011.654190

Wolverton, C. L., Lasiter, S., Duffy, J. R., Weaver, M. T., & McDaniel, A. M. (2018). Psychometric testing of the caring assessment tool: Administration (CAT-Adm). *SAGE Open Medicine*, *6*, 1–11. https://doi.org/10.1177/2050312118760739

Wright, B., & Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B., & Masters, G. (1982). *Rating scale analysis*. Mesa Press.

Wright, B., & Stone, M. (2004). *Making measures*. Phaneron Press.

Yeh, Y.-Y., & Wickens, C. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, *30*(1), 111–120.

VALIDITY AND TEST-LENGTH REDUCTION STRATEGIES

Appendix

Item Ordering of the *PSM6*

~~Table 9~~

Item Ordering From Rasch Analysis of Baseline PSM6 Performance

Question #	Domain	Measure	Model standard error
2	Geometry	3.36	0.47
14	Geometry	3.30	0.47
7	Geometry	2.80	0.38
13	Statistics and probability	2.62	0.36
10	Ratio and proportion	1.07	0.21
12	Ratio and proportion	–0.02	0.17
11	Number sense	–0.11	0.17
5	Number sense	–0.40	0.15
3	Number sense	–0.91	0.14
6	Expressions and equations	–0.98	0.14
9	Expressions and equations	–1.41	0.14
15	Ratio and proportion	–1.76	0.14
4	Ratio and proportion	–1.90	0.13
1	Statistics and probability	–2.80	0.13
8	Statistics and probability	–2.85	0.14