Elementary Mathematics Student Assessment

Measuring the Performance of Grade 1 and 2 Students in Counting, Word Problems, and Computation in Fall 2013

Robert C. Schoen
Mark LaVenia
Charity Bauduin
Kristy Farina

Suggested citation: Schoen, R. C., LaVenia, M., Bauduin, C., & Farina, K. (2016). *Elementary mathematics student assessment: Measuring the performance of grade 1 and 2 students in counting, word problems, and computation in fall 2013* (Research Report No. 2016-03). Tallahassee, FL: Learning Systems Institute, Florida State University. doi: 10.17125/fsu.1508170543

Elementary Mathematics Student Assessment (EMSA)

Measuring the Performance of Grade 1 and 2 Students in Counting, Word Problems, and Computation in Fall 2013

Research Report No. 2016-03

Robert C. Schoen

Mark LaVenia

Charity Bauduin

Kristy Farina

December 2016

Acknowledgements

A great many people were involved with the development, field-testing, data entry, data analysis, and reporting. Here we name some of the key players and briefly describe their roles, starting with the report coauthors.

Robert Schoen wrote the test and item specifications, developed items, coordinated the external review, edited and proofed the final forms, and managed the scoring and interpreting of results. Mark LaVenia performed the data analysis for the unidimensional item-response theory models, factor analytic models, reliability estimates, and regression models. Charity Bauduin managed the report-writing process. Kristy Farina served as the data manager and assisted with preparation of descriptive statistics for the present report.

We would like to acknowledge the reviewers of early drafts of the EMSA tests and express our gratitude for their contributions of expertise. These reviewers include Thomas Carpenter, Victoria Jacobs, Walter Secada, Juli Dixon, and Ian Whitacre.

Amanda Tazaz and Kristopher Childs managed the distribution and collection of tests and consent forms for students. Kristopher Childs and Juli Dixon managed the data entry and verification process.

Anne Thistle provided valuable assistance with editing the manuscript. Casey Yu provided valuable assistance with laying out the style and format of the final version of the report.

We are especially grateful to the Institute of Education Sciences at the U.S. Department of Education for their support and to the students, parents, principals, district leaders, and teachers who agreed to participate in the study and contribute to advancing knowledge in mathematics education. Without them, this work is not possible.

Table of Contents

List of Appendices

List of Tables

List of Figures

List of Equations

List of Abbreviations

CCSS-M .... Common Core State Standards for Mathematics

CDU .... Compare Difference Unknown

CFI.. Comparative Fit Index

CGI .... Cognitively Guided Instruction

DEA...Discovery Education Assessment

EMSA .... Elementary Mathematics Student Assessment

IRT .... Item Response Theory

ITBS.. Iowa Test of Basic Skills

JCU.. Join Change Unknown

JRU... Join Result Unknown

MG...Multiplication Grouping

MPAC...Mathematics Performance and Cognition

NA... Not Answered

PD ....Partitive Division

RMSEA ....Root Mean Square Error of Approximation

SCU .... Separate Change Unknown

SRU .... Separate Result Unknown

TIC ....Total Information Curve

TLI... Tucker-Lewis Index

UI...Unclear Intent

UIRT .... Unidimensional Item Response Theory

Executive Summary

The subject of this report is a pair of written, group-administered tests designed to measure the performance of grade 1 and grade 2 students at the beginning of the school year in the domain of number and operations. Because the tests are designed to be a measure of student achievement in elementary mathematics, we call them the Elementary Mathematics Student Assessment (EMSA) tests.

Purpose

The primary intended use of the EMSA tests was to serve as a covariate for students' baseline performance in statistical models estimating the impact of a teacher professional-development program on student achievement in mathematics as measured by the Iowa Test of Basic Skills (Dunbar et al., 2008) and the Mathematics Performance and Cognition (MPAC; Schoen et al., 2016). A secondary purpose was to serve as a test of baseline student achievement for the purpose of evaluating baseline equivalence of the students in schools assigned at random to treatment and control conditions.

This report is written for researchers and evaluators who may be interested in using the tests in the future or who wish to know about the psychometric properties of the tests.

Content

The contents of the EMSA tests are designed to align with core content in the *operations and algebraic thinking* and the *number and base ten* domains in the Common Core State Standards for Mathematics (CCSS-M) at grades 1 and 2, respectively (NGACBP & CCSSO, 2010). In a few instances, the content of the tests extends beyond the CCSS-M for the given grade level. These exceptions include *multiplication-grouping* word problems in grades 1 and 2 and a *partitive division* word problem in grade 2. The purpose of the focus on more advanced problems is to increase the ability of the test to discriminate among a wide range of levels of knowledge and understanding in the area of number and operations.

The final versions of the tests were the result of extensive development, feedback, and revisions from a variety of experts. The expert review verified the alignment of the content with the content of the CCSS-M at grades 1 and 2.

Test Specifications and Administration

The fall 2013 EMSA test has three main sections corresponding to counting and the number sequence, word problems, and computation. The test forms include 20 items at each grade level. Thirteen of the items are presented in a constructed-response format, and seven in a selected-response format.

On the basis of an iterative process of data modeling and item diagnostics, some of the items on the test forms were not used in the final scale. The final grade 1 scale uses data from 15 items. The final grade 2 scale uses data from 13 items. The two forms were not designed to be directly comparable.

Teachers administered the tests to their own students with the assistance of an administration guide and script (provided in Appendices C and D). Because of the paper-pencil format of the tests and the range in reading ability of the test takers, careful consideration was given to placement of the problems on each page and assisting students with identification of the correct page of the test during administration.

Sample and Setting

The 2013 EMSA tests were administered to 2,373 participating grade 1 and grade 2 students in 22 schools located in two public school districts in Florida during fall 2013. The school districts were implementing a curriculum based on the CCSS-M (NGACBP & CCSSO, 2010).

Scoring

Three first-order factors (Counting, Word Problems, and Computation) were regressed onto a single second-order factor (Math). The second-order total Math factor score is intended to serve as the overall achievement score on the pretest. Goodness-of-fit statistics varied but generally indicated that the specified measurement models provided a reasonable fit to the data. The Grade 1 model RMSEA statistic indicated mediocre fit, and the comparative fit index (CFI) and Tucker-Lewis index (TLI) statistics indicated reasonable fit: $\chi^2(87) = 1159.03$, $p < .001$; RMSEA = .10, 90% Confidence Interval (CI) [.10, .11]; CFI = .93; and TLI = .91. The Grade 2 model RMSEA statistic indicated reasonable fit, and the CFI and TLI statistics indicated close fit: $\chi^2(62) = 276.76$, $p < .001$; RMSEA = .06, 90% CI [.05, .06]; CFI = .96; and TLI = .95.

Reliability

The reliabilities of the test scales were determined on the basis of a composite reliability estimate for the higher-order Math factor and ordinal forms of Cronbach's α for the subscales. The grade 1 total Math composite reliability was .84; that for grade 2 was .89. On the grade 1 test, the α estimate for two of the three subscales exceeded or approximated the conventional target value of .8 (range .79 to .91). Grade 2 α estimates for all three subscales exceeded the conventional target value of .8 (range .82 to .86). The full research report presents diagnostic and supplementary analyses of scale reliability, including ordinal forms of Revelle's β and McDonald's ω_h coefficients and IRT information-based reliability estimates.

Concurrent and Predictive Validity

We examined evidence for the concurrent validity of the test by correlation of the test factor scores with the Discovery Education Assessment (DEA; DEA, 2010) scale scores. The DEA was used as an interim benchmark assessment by one of the participating districts in the sample. The correlations between the Math factor score and the DEA overall scale score were .69 in grade 1 and .61 in grade 2; both correlations were statistically significant at $p < .001$. The statistically significant, moderately-sized correlation coefficients provide some, albeit modest, evidence of concurrent validity for the test as it relates to the DEA district-administered interim assessment.

Evidence for the predictive validity of the test was examined by regression of the standard scores for the level 7 and level 8 Iowa Test of Basic Skills (ITBS; Dunbar et al., 2008) tests on the fall 2013 EMSA Math factor scores for grades 1 and 2, respectively. Regression results suggested that the fall 2013 EMSA Math score was a moderate to strong predictor of students' scores on the ITBS Math Problems test, where an $R^2_{Adjusted}$ of .41 was found for grade 1 and an $R^2_{Adjusted}$ of .49 was found for grade 2. The EMSA Math scores provided more modest predictive power with the ITBS Math Computation test, where an $R^2_{Adjusted}$ of .23 was found for grade 1 and an $R^2_{Adjusted}$ of .30 was found for grade 2. All of these relations were statistically significant at $p < .001$. The regression analyses suggest the EMSA to be an appropriate student mathematics achievement covariate in analyses that use the ITBS tests as outcomes, where the

results suggest the test is particularly well suited for this purpose in analyses using scores from the ITBS Math Problems test as the outcome variable.

Summary

We report on the initial validation efforts examining the substantive, structural, and external validity (Flake, Pek, & Hehman, 2017) for the fall 2013 EMSA tests. These tests were designed to be a measure of student achievement in grades 1 and 2 for use as a student pretest covariate in the study of the effects of a mathematics-teacher professional-development program in mathematics. EMSA test items were constructed and reviewed by mathematicians and mathematics education experts and measure student achievement in the domain of operations and algebraic thinking as well as number and base ten. The development process, model fit, and scale-reliability estimates meet the basic standards for educational measurement. Test scores are moderately correlated with the scores of policy-relevant, standardized tests used to measure student achievement in grades 1 and 2. The EMSA tests appear to be sufficiently well suited for their primary intended use as a test covariate for the evaluation of educational interventions involving grade 1 and grade 2 students.

1. Introduction and Overview

The fall 2013 EMSA tests were designed to measure student mathematics performance at the beginning of grade 1 and grade 2. The items focus on tasks involving counting, word problems, and computational problems.

The test-development process involved multiple iterations of item and test blueprint development, review of items and the test blueprint by experts in mathematics and mathematics education, and extensive revisions and proofreading of the items, sequence, and formatting. Experts provided feedback on the accuracy of the mathematics content, clarity of questions, number choices in the selected-response items, overall length of the test, and predictions about how students could potentially misinterpret the items in ways that might obscure their ability to measure student knowledge and ability. Experts also reviewed the items on both tests to determine the extent of the alignment of the items with the domains of counting and algebraic thinking in the CCSS-M (NGACBP & CCSSO, 2010).

The EMSA tests were designed to be administered in a whole-group setting in a paper-pencil format. The students' classroom teachers were asked to administer the tests during the first two weeks of the school year. The teachers were given an administration guide explaining how to administer the tests and a script to use while administering them. Questions were read aloud to students, and students either filled in a box with the correct number for open-ended items or shaded bubbles to indicate their responses to multiple-choice items. Teachers were encouraged to allow students to use manipulatives in accordance with their typical classroom practice.

The immediate purpose of the tests was for use as a student pretest covariate in a randomized controlled trial evaluating the impact of a teacher professional-development program on student achievement in the domains of number, operations, and algebraic thinking. In the state and school districts where the efficacy trial took place, no uniform measure of student mathematics achievement was used with kindergarten, grade 1, or grade 2 students. A measure of student achievement in mathematics was desired for the purposes of investigating baseline equivalence of participating schools and as a student-level covariate in statistical models estimating the impact of the program on student achievement.

1.1. Test Overview

The EMSA tests contain 20 items on each grade level test. These items are grouped into three sections for the administration of the tests: Counting, Word Problems, and Computation. Table 1 provides a listing of the sections and number of items administered to grade 1 and grade 2 students.

Table 1. Number of Items That Remained on the Fall 2013 Tests After Screening and Respecification

Section	Grade 1	Grade 2	Common items
Counting	3	3	0
Word Problems	7	7	0
Computation	10	10	3
Total	20	20	3

Although the two tests consist of the same three sections and approximately the same number of items, they are not designed to be vertically scaled. Only three of the items on the two tests are identical, and all three of those are in the Computation section. When individual items on the grade 1 and grade 2

tests are similar (but not identical), the questions on the grade 2 test involve higher numbers so as to increase the difficulty proportionally with age and to elicit information about how these older students make sense of operations on multidigit whole numbers.

1.1.1. Section 1: Counting

The initial section of the test was intended to ask students questions about number and quantity. Table 2 shows the number of items and the question asked within each item. All three of the items in the Counting section for both the grade 1 and grade 2 tests have a constructed-response format.

Table 2. Items in the Counting Section

Grade 1 test item number	Grade 1 item	Grade 2 test item number	Grade 2 item
1[a]		1	
2		2	
3		3	

As Table 2 demonstrates, two of the grade 1 items in the Counting section are identical in structure to two of the grade 2 items, but the grade 2 items involve higher numbers, for two main reasons. The numbers in the beginning-of-year grade 1 test are less than 20 to align with expectations in the state mathematics curriculum standards (and the CCSS-M). Two-digit numbers are used in the grade 2 test items as a means of increasing difficulty of items. This increase was used as a strategy to improve the ability of the test to discriminate among students with different ability levels and to improve alignment with the learning expectations in the curriculum standards.

1.1.2. Section 2: Word Problems

The second section of the test contains a set of word problems representing a range of difficulty. Table 3 provides the sequence of word problems in this section. For brevity, the list indicates only the type of problem and the numbers presented in the problem. All the Word Problems items in both tests used a selected-response (i.e., multiple-choice) format. This format is consistent with the format of the ITBS tests (Dunbar et al., 2008). The ITBS tests comprise two of the three outcomes of interest in the randomized controlled trial in which the fall 2013 EMSA data were used as a student achievement covariate.

Table 3 shows that both the grade 1 and grade 2 tests included *join result unknown* (JRU), *join change unknown* (JCU), *separate result unknown* (SRU), and *multiplication grouping* (MG) problems. Although the grade 1 and 2 tests contain problems of the same problem types, the wording, contexts, and number choices on the two tests differ. The numbers on the grade 2 test were selected with the intent to increase the difficulty level of the item for use with the grade 2 population.

Table 3. Summary of Items Used in the Word Problems Section

Grade 1 test item number	Grade 1 item	Grade 2 test item number	Grade 2 item
4		4	
5		5	
6		6	
7		7	
8		8	
9		9	
10		10	

Note. See the list of the abbreviations for elaboration on the problem type categories (Carpenter et al., 1999).

1.1.3. Section 3: Computation

The Computation section includes items asking students to perform calculations involving addition and subtraction on whole numbers. Table 4 presents the sequence of problems in the Computation section of the tests. Three computation items on the grade 1 and grade 2 tests are identical: evaluation of ███ , ███ , and ███ .

Table 4. Items in the Computation Section

Grade 1 test item number	Grade 1 item	Grade 2 test item number	Grade 2 item
11		11	
12		12	
13		13	
14		14	
15		15	
16		16	
17		17	
18		18	
19		19	
20		20	

1.2. Administration of Test

Tests were delivered to schools by project staff during the week of preplanning (i.e., the week before students returned to school for the year). Teachers were given detailed instructions on how to administer the tests. The tests were accompanied by a document for teachers—provided here in Appendices C and D—containing detailed test-administration instructions, including a script to use while administering the tests.

Teachers were asked to write the students' names on the front covers of the tests to increase legibility and accuracy in data entry. Teachers were also instructed to permit students to use manipulable materials if that was common practice in their classrooms. For the first two sections of the test, teachers were instructed to read the problems aloud to students—in their entirety—to reduce the effect of reading ability on students' mathematics performance. Reading problems aloud to students is consistent

with the administration procedures for the ITBS and the Mathematics Performance and Cognition (MPAC) interview, the two outcome measures used for the randomized controlled trial. As necessary, teachers were encouraged to provide appropriate testing accommodations for students in accordance with their individual educational plans. Teachers were instructed to insert completed tests into an opaque, sealed envelope and deliver the envelopes to the front office for project personnel to pick up during a window of time outlined in the administration instructions.

We acknowledge that teacher administration presents the potential for breaches in security. These were not high-stakes tests, so strict security was not a high priority. In this case, teachers and schools were trusted to administer the tests in accordance with the instructions.

1.3. Description of the Sample

The student sample included 2,373 students (1,226 grade 1 and 1,147 grade 2) with consent to participate. The student sample came from the classrooms of participating grade 1 and 2 teachers representing 22 schools in two diverse public school districts (7 schools in one district; 15 in the other) in Florida. Grade 1 and 2 teachers in these schools elected to participate in a large-scale, cluster-randomized controlled trial evaluating the efficacy of a teacher professional-development program in mathematics. Half of the schools in this sample were assigned at random to the treatment condition; the other half to the control condition. Our sampling procedure attempted to measure all grade 1 and grade 2 students in participating teachers' classrooms. Other than the requirement for parental consent in order for data on students to be collected, no exclusion criteria were applied that would have limited the sample by student characteristic. Table 5 presents the student demographics for the total participating student sample as of fall 2013 and the subsample of students for whom fall 2013 measurement with the EMSA was conducted.

Table 5. Student Sample Demographics

Characteristic	Total student sample (*n* = 2,631)		Student test sample (*n* = 2,373)	
	Proportion	*n*	Proportion	*n*
Gender				
Male	.48	1,261	.48	1,144
Female	.47	1,247	.48	1,135
Unreported	.05	123	.04	94
Grade				
1	.50	1,326	.51	1,226
2	.50	1,305	.49	1,147
Race/Ethnicity				
Asian	.04	115	.05	108
Black	.17	459	.18	416
White	.35	912	.36	852
Other	.03	70	.03	65
Hispanic	.35	910	.34	814
Unreported	.06	165	.05	118
English language learners	.21	553	.21	498
Eligible for free or reduced-price lunch	.58	1,523	.58	1,364
Exceptionality				
Students with disabilities	.07	184	.07	166
Gifted	.04	97	.04	91
Unknown	.06	165	.05	118

Note. Proportion provided reflects percentage of total sample. Some characteristic categories are not mutually exclusive. Students with unreported demographic information are represented in the "Unknown" category. The Asian, Black, and White categories are non-Hispanic.

2. Test Development

2.1. Content

The content standards at grades 1 and 2 in the CCSS-M (NGACPB & CCSSO, 2010) were used to provide guidelines for content specifications. Overall, the focus of the test is on number and operations, but it includes some items designed to favor students who have a solid grasp of place-value concepts. The numbers used on the test are limited to positive integers (i.e., Counting numbers) between 1 and 100. Computation items presented symbolically involve applying the addition or the subtraction operation with exactly two positive integers. Problems involving subtraction result in a difference with a positive, integer value. Word problems involve additive situations as well as grouping situations that could be solved by multiplication, division, addition, counting strategies, or direct place-value understanding (Carpenter et al., 1999).

2.2. Test Specifications

Test design involved finding an optimum point at the intersection of three potentially competing goals: (1) sample a range of difficulty of problems and cognitive demand to reflect the focus of the teacher professional-development program goals and the learning goals outlined in grades 1 and 2 in the CCSS-M, (2) serve as a reasonably strong student-level test covariate to explain some of the variance in the ITBS and MPAC interview data, and (3) minimize the test-taking burden on teachers and students.

The Counting and Word Problems sections of the test include only one item per page to minimize student distraction and confusion. Rather than using Arabic numerals as page numbers or to enumerate items, we used a child-friendly image to identify each page. We used graphics in order to be as considerate as possible of the test taker (who may not read Arabic numerals fluently). Figure 1 provides one example of these graphics.

Figure 1. One of the images used in place of a page number.

Beginning-of-year grade 1 students, in particular, may not recall all of their numerals, and numbered pages could cause confusion and anxiety. The large and easily distinguished image is also useful for the test administrator as a way to verify from across the room that all students have turned to the correct page. Moreover, the ITBS test forms use a similar tactic, so this test serves as practice for that type of format.

Response types include selected-response (i.e., multiple-choice) and constructed-response items. All of the constructed-response items are short answer; none of them requires extended or elaborated responses. Sample items with examples of responses are provided on the first page of the test for the administrator to demonstrate how students are expected to respond (e.g., completely shade the bubble, write a numeral in a rectangular area designated for the response).

Selected-response options are ordered from least to greatest and from left to right. Bubbles are centered beneath each response option, and responses are centered horizontally across the page. Test items were reviewed internally for bias and sensitivity in an effort to neutralize any need for vocabulary

development with students. Whenever possible, word problems are written to avoid the use of so-called keywords (i.e., altogether, in all, left).

Although the tests designed for the two grade levels have the same three sections (i.e., Counting, Word Problems, Computation), the tests are not designed to be vertically scaled or equated. The grade 2 test was designed to be more difficult than the grade 1 test.

2.3. Item Development

The items were written by the first author of the present report. Schoen holds postsecondary degrees in atmospheric science, mathematics, and mathematics education. He has extensive experience developing assessment items and scales designed to measure student cognition and achievement in early elementary mathematics as well as teacher knowledge and beliefs. The items were reviewed by other individuals with expertise in elementary education, assessment, and mathematics.

The development process for the tests consisted of several phases. These phases included:

1. Analysis of the goals of the mathematics professional-development program we were evaluating: Cognitively Guided Instruction (CGI).
2. Review of the learning goals delineated in the CCSS-M grades 1 and 2.
3. Review of literature and related measures in the domain of number and operations at grades 1 and 2.
4. Creation of a draft test blueprint.
5. Review of item and scale performance from the 2013 version of the test; review of student responses for those items used on the 2013 tests.
6. Development of a first written draft of the grade 1 and grade 2 test items.
7. Internal review of drafted tests by members of the research team as well as review by several members of the project advisory board.
8. Revision of drafts based upon feedback.

Because the tests were used in the evaluation of a program related to CGI, an extensive body of literature related to CGI was reviewed carefully (cf. Carpenter et al., 1989, 1999; Fennema et al., 1996; Jacobs et al., 2007). The CGI program is focused on number (including place value), operations, and algebraic thinking. As part of a strategy to avoid overalignment with the intervention, we also completed a review of the learning goals set forth in the CCSS-M (NGACBP & CCSSO, 2010). The topics at the intersection of the program goals and the expectations outline in the CCSS-M provided the starting place for defining the content of the test.

Once the blueprint was developed, a draft set of items was written and reviewed internally by the research team, which consists of experts in mathematics, mathematics education, educational psychology related to student thinking in mathematics, and educational measurement. After this internal review, the draft set of items and testing format were revised and sent to advisory board members Thomas Carpenter, Victoria Jacobs, and Ian Whitacre for review and feedback. Dr. Carpenter provided extensive feedback based on his experience assessing students, and the items were heavily revised on the basis of his recommendations. Revised versions of the items were then internally reviewed by personnel working on the larger study.

2.4. Test Design and Assembly

The student tests consist of three sections: Counting, Word Problems, and Computation. The Counting section consists of three items aimed at measuring students' understanding in the domain of counting and cardinality. All of the Counting items use a constructed-response format, in which the students are expected to write each answer as a numeral in a designated box. The Word Problems section includes seven items, all of which use a selected-response format and offer five response options for each item. The response options are always numerals and are ordered from least to greatest, from left to right. The students are directed to fill in the circles below their answer choices. The Computation section consists of 10 items presented as open equations. Each problem is presented as a single equation involving either the addition or the subtraction operator and exactly two numerals. Each is presented in the standard (i.e., $a + b = c$, $a − b = c$) form (Stigler et al., 1986; Schoen et al., manuscript under review) with an open box providing a place for the student to write the numeral representing the sum or difference.

In the Counting and Word Problems sections, only one problem is displayed per page so that students will not record their answers in the wrong places or be overwhelmed by too much text on the page. Computation items are presented with multiple items split across two pages. In an effort to avoid confusion, as well as to match the format of the ITBS outcome measure, a line is placed after each Computation item on the page. The grammar used in word problems was reviewed by those with experience in teaching emergent bilingual students. The font used in the final version of the test is large (18-point) to increase legibility. Copies of the grade 1 and grade 2 tests are presented in Appendices A and B, respectively.

2.5. Test Production and Administration

The tests, administration guides, and consent forms were printed at the university and distributed to the participating schools. Tests were printed single-sided on 20-pound, white paper in the 18-point Calibri font.

Administration guides were designed and created for teachers to use while administering the tests. They provide an overview of the tests, describe the administration process and directions, explain how to submit completed tests, and provide a full script to be read verbatim during administration of the test. In addition, the administration guides include a student information sheet on the last page. Teachers completed this sheet to provide student and class information (e.g., student names, student ID numbers, testing accommodations provided) and returned it with the completed student tests. The administration guide was repeatedly reviewed, edited, and proofread by research project staff before the final version was produced. The final forms of the test administration guides for grades 1 and 2 are presented in Appendices C and D, respectively.

Participating teachers were provided with a test packet containing:

- Testing administration guide (for the corresponding grade level)
- Class set of student tests
- Parental consent forms
- Student information sheet

These materials were distributed to the teachers participating in the study through the main office personnel or principal-appointed designee. Test materials were distributed to the main offices at school sites on August 5–9, 2013. Teachers were instructed to administer the tests during the first three weeks of school.

Test administrators (which were usually the participating teachers) were directed to read each math problem aloud to students in accordance with the administration script. In addition, they were asked to provide and allow students to use manipulatives, like counters or linking cubes, during the test. If students generally had testing accommodations as a result of IEP, ELL or 504 plans, then the teacher was asked to provide any and all required accommodations for those individual students and to document the accommodation on the student information sheet. The test is not timed, so test administrators were instructed to allow students adequate time to answer all of the questions.

Upon conclusion of administration, teachers were instructed to submit all testing materials (i.e., test administration guide, student test booklets, student information sheet, student booklist form, and parental consent forms) to their principals or designees. Teachers were asked to return only test booklets completed by those students with corresponding signed parental consent on the parental consent form. The principal or designee placed the testing materials in the main office at the front desk for pickup. Members of the project team picked up test materials during the last two weeks of September 2013.

Teachers who presented extenuating circumstances to the research team and did not administer the test during the administration window or missed the materials pickup date were handled on a case-by-case basis with respect to when to administer the test and arrangement of a materials pickup date. Very few instances of these special cases arose.

3. Data Entry and Analysis Procedures

3.1. Data Entry and Verification Procedures

Research assistants typed student responses into an Excel spreadsheet with response fields validated to allow only whole numbers and accepted codes for missing items. Missing responses were coded in two ways: "UI" indicated Unclear Intent, and "NA" indicated Not Answered. Research assistants were given the task of interpreting both the student's handwriting and the student's intent, with the goal of entering the student's intended response exactly as it was written. Because this assessment was administered to grade 1 students at the beginning of the school year, many student responses displayed immature handwriting that took careful consideration. As a result, the assistants met regularly to discuss, and come to agreement on, student responses. In most cases, the discussion was over which numerals the student wrote, although on occasion discussions to determine which of the numerals a student wrote were intended as the answer. The UI code was used when the committee could not come to an agreement about the student's intended response or when the student's response was too far from standard numeric representations to be interpreted. Common examples of responses that required interpretation and discussion are listed below, with a description of the decision that was made.

- *The answer was "7", but the student wrote "07" on the answer line.* Correct responses preceded by a zero were interpreted as correct. In this example, the exact student response would be entered as written.
- *The answer was "13", but the student wrote "31" on the answer line.* Numeric reversals were entered as written, and interpreted as incorrect. Committee members agreed that although students who responded "31" may have intended to write "13," evidence was insufficient to support that claim.
- *The answer was "3" and the student wrote a backwards three.* Backwards numerals were interpreted as though they were written correctly. No indication was made during data entry to signal that a numeral was written backwards. This decision only applies to individual digits, and did not override the decision for reversals of multidigit numbers.

Many items brought to committee for review were flagged by the research assistant as difficult to interpret. To ensure data quality, a sample of 10% of the data was randomly selected for review. These data were entered by a second reviewer and compared to the original entries. The two entries were compared for agreement on response given for each item to confirm that agreement was within an acceptable range. Once both entries were scored as correct or incorrect for all items, the overall agreement between the two was 99%.

3.2. Data Analysis

All analyses were performed in Mplus version 7.11 (Muthén & Muthén, 1998-2012), with the exception of the estimation of Cronbach's α, Revelle's β, and McDonald's ω_h hierarchical reliability coefficients, which were performed in R 3.1.2 (R Development Core Team, 2014) using the psych package (Revelle, 2016) α, splithalf, ω_h, and polychoric functions.

Our investigation consisted of five steps. We aimed (1) to screen out items that demonstrated outlier parameter estimates when fit to a unidimensional framework, (2) to evaluate item performance structured in accordance with the three-factor blueprint and drop items that demonstrate low salience

with their respective factor, (3) to respecify the structure of the model from one of correlated factors to one of a single second-order factor and three first-order factors, (4) to estimate reliabilities for the test overall and for each subscale, and (5) to estimate the concurrent and predictive validity of the test for each grade level.

The first step was to screen the initial set of items within a 2-parameter logistic (2-pl) unidimensional item response theory (UIRT) framework. Discrimination and difficulty parameters were inspected. An item was flagged for removal if (a) its discrimination estimate was less than .4 or greater than 3 or (b) the absolute value of its difficulty estimate was greater than 3. These cut points were not strictly enforced. For example, items with low discrimination that appeared to fill a void along the difficulty continuum received special consideration for being retained.

The second step was to fit the screened data to a correlated-trait item-factor analysis (confirmatory factor analysis with ordered categorical indicators) model that paralleled a 3-factor model structure specified by the principal investigator in consultation with item reviewers.

We used the model chi-square (χ^2), RMSEA, CFI, and TLI to evaluate overall model fit. Following guidelines in the structural-equation modeling literature (Browne & Cudeck, 1992; MacCallum, Browne, & Sugawara, 1996), we interpreted RMSEA values of .05, .08, and .10, as thresholds of close, reasonable, and mediocre model fit, respectively, and interpreted values > .10 to indicate poor model fit. Drawing from findings and observations noted in the literature (Bentler & Bonett, 1980; Hu & Bentler, 1999), we interpreted CFI and TLI values of .95 and .90 as thresholds of close and reasonable fit, respectively, and interpreted values < .90 to indicate poor model fit. We note that little is known about the behavior of these indices when they are based on models fit to categorical data (Nye & Drasgow, 2011), which adds to the chorus of cautions associated with using universal cutoff values to determine model adequacy (e.g., Chen, Curran, Bollen, Kirby, & Paxton, 2008; Marsh, Hau, & Wen, 2004). Because fit indices were not used within any of the decision rules, a cautious application of these threshold interpretations bears on the evaluation of the final models but has no bearing on the process employed in specifying the models.

Confirmatory factor analysis models with standardized factor loadings > .7 in absolute value are optimal, as they ensure that at least 50% of the variance in responses is explained by the specified latent trait. In practice, however, this criterion is often difficult to attain while maintaining the content representativeness intended for many scales. Researchers working with applied measurement (e.g., Reise, Horan, & Blanchard, 2011) have used standardized factor loadings as low as .5 in absolute value as a threshold for item salience. In accordance with this practice, we aimed to retain only items in the final model that had standardized factor loading estimates > .5 and unstandardized factor loading *p*-values < .05.

The third step was to respecify the reduced set of items with a higher-order factor structure, in which the three first-order factors were regressed onto a single second-order factor. The purpose of respecifying the factor structure as a higher-order model was to select a more parsimonious factor structure that provided the pragmatic benefit and utility of having a single underlying factor (and composite score).

The fourth step was to inspect the scale reliabilities, which we did by calculating the composite reliability for the higher-order total Math factor and estimating ordinal forms of Cronbach's α, Revelle's β, and McDonald's ω_h for the subscales. As a supplementary analysis, we also estimated the reliability for the total Math scale, except modeled as a single factor on which the reduced set of items loaded directly. To

evaluate reliability coefficients, we applied the conventional values of .7 and .8 as the minimum and target values for scale reliability, respectively (Nunnally & Bernstein, 1994; Streiner, 2003).

Using the equation described by Geldhof, Preacher, and Zyphur (2014), we calculated the composite reliability as the squared sum of unstandardized second-order factor loadings divided by the squared sum of unstandardized second-order factor loadings plus the sum of the first-order factor residual variances. The first-order factors are Counting, Word Problems, and Computation. Equation 1 shows the equation for the composite reliability for the second-order Math factor, where λ is the unstandardized second-order factor loading and ζ is the residual variance for the respective first-order factor.

$$\text{Composite reliability} = \frac{(\lambda_{CNT} + \lambda_{WP} + \lambda_{CMP})^2}{(\lambda_{CNT} + \lambda_{WP} + \lambda_{CMP})^2 + (\zeta_{CNT} + \zeta_{WP} + \zeta_{CMP})} \tag{1}$$

This calculation is analogous to the classical conceptualization of reliability as the ratio of true score variance to the true score variance plus error variance.

For our estimation of ordinal forms of Cronbach's α, Revelle's β, and McDonald's ω_h, we executed the procedure described by Gadermann, Guhn and Zumbo (2012). Cronbach's α is mathematically equivalent to the mean of all possible split half reliabilities and Revelle's β is the worst split half reliability. Only when essential τ equivalence (i.e., unidimensionality and equality of factor loadings) is achieved will α equal β; otherwise, α will always be greater than β. Variability in factor loadings can be attributable to microstructures (multidimensionality) in the data: what Revelle (1979) termed *lumpiness*. McDonald's ω_h models lumpiness in the data through a bifactor structure. The relation between α and ω_h is more dynamic than that between α and β, as α can be greater than, equal to, or less than ω_h, as a result of the particular combination of scale dimensionality and factor loading variability. We investigated these scale properties by examining the relation among coefficients α, β, and ω_h through the four-type heuristic proposed by Zinbarg, Revelle, Yovel, and Li (2005).

The reduced set of items in the final model of the test were fit to a 2-pl UIRT model to produce a total information curve (TIC) for each grade-level test for the purpose of judging scale reliability across the distribution of person ability. Inspecting the TICs allowed us to make the conversion from information function to reliability along a given range of person abilities with Equation 2.

$$\text{Reliability} = \frac{\text{Information}}{\text{Information}+1} \tag{2}$$

Accordingly, information of 2.33 converts to reliability of approximately .70 and information of 4.00 converts to a reliability of .80, for example. Equation 2 derives from the classical test theory equation of reliability = true variance / (true variance + error variance). Applied to an IRT framework, where error variance = 1 / information, the equation works out to reliability = 1 / 1 + (1 / information), which coverts algebraically to information / (information + 1) (http://www.lesahoffman.com; cf. Embretson & Reise, 2000).

The reliability estimates directly relevant to the scales as described and presented as the final models in this research report are the composite reliability for the higher-order Math factor and the α, β, and ω_h reliability coefficients for the subscales. That is, the α, β, and ω_h reliability coefficients and the 2-pl UIRT information-based reliability estimates for the total Math scale apply to structures and modeling approaches different from those of the higher-order structure described in this research report. These supplementary analyses of reliability for the total Math scale were conducted as part of our endeavor to obtain a broad understanding of how the items from the final model worked together and are presented principally with the purpose of thoroughness and transparency in reporting.

The fifth, and final, step of our investigation of the tests' psychometric properties was to inspect for evidence of concurrent and predictive validity for the scales. All analyses of concurrent and predictive validity involved first saving the factor scores from the final higher-order factor model for the grade 1 and grade 2 tests; then, as manifest variables, the factor scores were merged into a file containing criterion-relevant scores to which the tests were compared. The criterion for the concurrent validity analyses was the DEA (DEA, 2010). For the predictive validity analyses, the criterion was the ITBS Math Problems test and ITBS Math Computation test (Dunbar et al., 2008).

We investigated evidence of concurrent validity of the tests by correlating the tests' factor scores with scores from the DEA. The DEA was used by one of the participating districts (District 2) in the current study as an interim benchmark assessment across three time points annually. District 2 provided the DEA data for all consenting students. For the investigation of concurrent validity, we used the fall 2013 administration of the DEA, which had an assessment window of August 19 through October 4, 2013. Teachers were instructed to complete administration of the EMSA tests between August 17 and August 30, 2013. Some teachers were granted an extension to administer the test as late as September 30, 2013. Additional time was granted on an as-needed basis. The DEA data comprise an overall scale score and total number correct for each of three subdomains: Operations, Base Ten, and Measurement and Data. Correlations were estimated between the test factor scores and the DEA total and subdomain scores. Correlation coefficients and corresponding p-values are reported, and correlations > .7 are interpreted to indicate scale correspondence.

We investigated evidence of predictive validity by regressing the ITBS tests' standard scores onto the grade 1 and grade 2 tests' factor scores. Standardized beta (β) coefficients, corresponding p-values, and adjusted R-squared ($R^2_{Adjusted}$) coefficients of determination are reported, and an $R^2_{Adjusted}$ > .4 is interpreted to indicate that a substantial proportion of variance in the target outcome was explained by the test score. The ITBS tests were administered to the sample spring 2014. For the predictive validity analyses, the sample was constrained to the control group students only.

4. Results

The following sections describe the process of item screening, evaluation, and model respecification that was used to determine the final set of items. Before we report on the detailed results of those analyses, we provide a blueprint for the final tests in section 4.1 that shows the number of items corresponding to the three lower-order factors in the final scale for the tests. After providing the blueprint, we proceed chronologically through the steps of screening, model specification, and evaluation.

4.1. Three-factor Test Blueprint

Table 1 in section 1.1 provided an overview of the original items offered to students on the 2013 EMSA. Initially the grade 1 test included 20 items, as did the grade 2 test. Some of the items were dropped from the scales because of poor item statistics. Table 6 provides an overview of the number of items that remained in the final scales for grades 1 and 2.

Table 6. Number of Items That Remained on the Fall 2013 Tests After Screening and Respecification

Section	Grade 1	Grade 2	Common items
Counting	2	3	0
Word Problems	4	4	0
Computation	9	6	2
Total	15	13	2

4.2. Item Screening

Tables 7 and 8 present the full set of items on the grade 1 and grade 2 student tests, respectively. The tables report the proportion answered correctly as well as the 2-pl UIRT discrimination and difficulty parameter estimates for each item on each test. For ease of reference, we presented in italics the entries for items that remained in the final model after undergoing the full procedure of screening, evaluation, and respecification. Also for ease of reference, we have inserted a column that names which section each item belonged to, according to the item blueprint. Tables 7 and 8 present the items in the order administered and organizes them according to whether the item structure was that of counting, word problem, or computation prompt. Interested readers will find information about the most common incorrect responses to each item in Appendix E.

4.2.1. Grade 1 Test Item Screening

Table 7 reveals that, on the grade 1 test, the absolute value of the difficulty estimate item for item 1 exceeded the maximum acceptable value for item difficulty. The high proportions correct observed for item 1 (.97) is consistent with the outlier estimate for its difficulty parameter.

Table 7. Grade 1 Test Item Descriptions, Percentage Correct, and Unidimensional IRT Parameters

Section	Item description	Proportion correct	2-pl UIRT parameters	
			Discrimination	Difficulty
Counting				
Item 1[a]		*0.967*	*0.278*	*−7.331*
Item 2		0.764	0.657	−1.296
Item 3				
		0.540	0.821	−0.156
Word Problems				
Item 4[a]		*0.786*	*0.523*	*−1.691*
Item 5		0.338	0.716	0.706
Item 6[a]		*0.403*	*0.403*	*0.634*
Item 7		0.171	0.589	1.863
Item 8[a]		*0.576*	*0.591*	*−0.367*
Item 9		0.350	0.832	0.597
Item 10		0.363	0.675	0.616
Computation				
Item 11		0.668	0.896	−0.646
Item 12		0.389	1.694	0.321
Item 13		0.794	0.806	−1.307
Item 14		0.316	1.377	0.587
Item 15		0.785	0.753	−1.304
Item 16		0.338	1.494	0.497
Item 17		0.281	1.268	0.738
Item 18		0.254	1.522	0.792
Item 19[a]		*0.396*	*0.576*	*0.519*
Item 20		0.613	0.814	−0.448

Note. n= 1,226 grade 1 students who completed the EMSA in fall 2013. 2-pl UIRT refers to 2-parameter logistic unidimensional item response theory model. Discrimination estimates use a 1.702 scaling constant to minimize the maximum difference between the normal and logistic distribution functions (Camilli, 1994).
[a]Entries for items that were removed during the calibration process and not used in the final scale is presented in italics.

We plotted the discrimination and difficulty parameters to inform our decision on retaining or dropping items. Figure 2 presents the grade 1 difficulty-versus-discrimination scatterplot. Because several satisfactorily discriminating items were included near the lower end of the difficulty range, the lower-end of the difficulty distribution seemed to be adequately represented without the retention of item 1. We therefore determined item 1 not to pass the item screening.
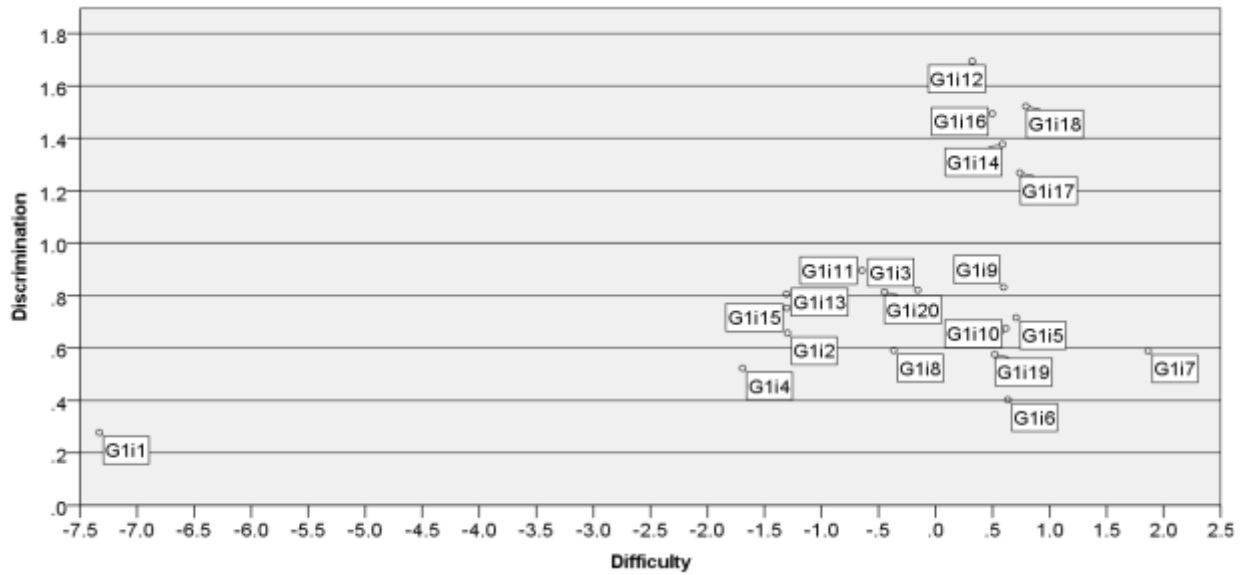
Figure 2. Grade 1 test 2-pl unidimensional item response theory (UIRT) difficulty-vs.-discrimination scatterplot.

4.2.2. Grade 2 Test Item Screening

Table 8 reveals that no items on the grade 2 test have outlier discrimination or difficulty estimates. Accordingly, all items on the grade 2 test were determined to pass the item screening.

Table 8. Grade 2 Test Item Descriptions, Descriptive Statistics, and Unidimensional IRT Parameters

Section	Item description	Proportion correct	2-pl UIRT parameters Discrimination	Difficulty
Counting				
Item 1[a]		0.888	0.875	−1.871
Item 2		0.728	0.968	−0.859
Item 3		0.679	1.012	−0.638
Word Problems				
Item 4[a]		*0.892*	*0.677*	*−2.236*
Item 5		0.531	1.089	−0.090
Item 6		0.558	0.785	−0.222
Item 7		0.740	1.103	−0.856
Item 8[a]		*0.742*	*0.564*	*−1.296*
Item 9		0.491	0.748	0.048
Item 10[a]		*0.667*	*0.731*	*−0.713*
Computation				
Item 11[a]		*0.922*	*0.718*	*−2.500*
Item 12		0.822	0.548	−2.909
Item 13		0.840	0.903	−1.492
Item 14[a]		*0.762*	*0.718*	*−1.208*
Item 15[a]		*0.676*	*0.634*	*−0.829*
Item 16		0.658	0.826	−0.623
Item 17		0.635	0.682	−0.592
Item 18		0.591	1.041	−0.303
Item 19[a]		*0.425*	*0.595*	*0.366*
Item 20		0.532	0.726	−0.124

Note. n= 1,147 grade 2 students who completed the EMSA in fall 2013. 2-pl UIRT refers to 2-parameter logistic unidimensional item response theory model. Discrimination estimates use a 1.702 scaling constant to minimize the maximum difference between the normal and logistic distribution functions (Camilli, 1994).
[a]Entries for items that were removed during the calibration process and not used in the final scale is presented in italics.

We plotted the discrimination and difficulty parameters to inform our decision on retaining or dropping items. Figure 3 presents the grade 2 difficulty-versus-discrimination scatterplot.
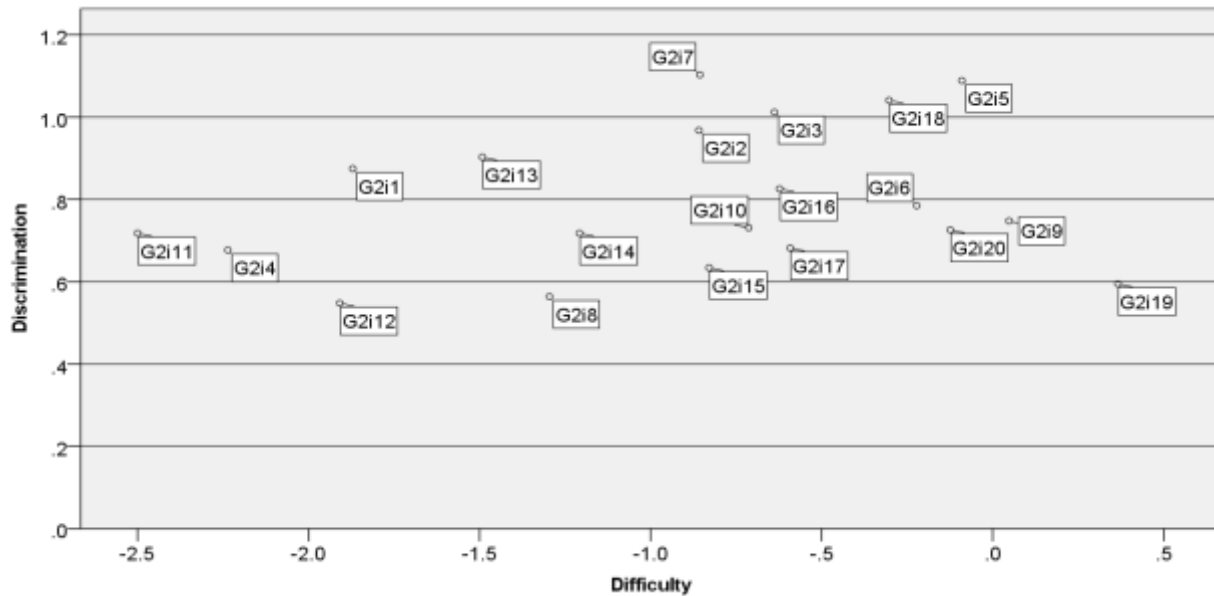
Figure 3. Grade 2 test 2-pl UIRT difficulty-vs.-discrimination scatterplot.

4.3. Correlated-Trait Model Evaluation

4.3.1. Grade 1 Correlated-Trait Model Evaluation

The initial grade 1 correlated-trait model contained all items that were administered on the grade 1 test except item 1. All items in the initial model had statistically significant unstandardized factor loading ($p <$ .001). Four items (4, 6, 8, and 19) had standardized factor loadings near the factor-loading minimum acceptable value of .5. Upon inspection of the standardized loadings for items 4 (.50), 6 (.52), 8 (.62), and 19 (.56) and their representation of the range of item difficulty, as well as consideration of their relative contribution toward the content validity of the scale, we decided that all four items could be dropped for the revised model.

We then fit the data for the reduced set of grade 1 items to a revised correlated-trait structure and evaluated the factorial validity of the model on the basis of overall goodness of fit and interpretability, size, and statistical significance of the parameter estimates. The revised grade 1 correlated-trait model fit statistics indicated mediocre fit by the RMSEA statistic and reasonable fit by the CFI and TLI statistics: $\chi^2(87) = 1159.026$, $p <$ .001; RMSEA = .100, 90% CI [.095, .105]; CFI = .929; and TLI = .914. All unstandardized factor loadings for the revised grade 1 model were statistically significant. Table 9 presents the standardized factor loadings for the initial and revised correlated-trait model. All standardized factor loadings for the revised grade 1 model were above the minimum acceptable value of .5, and most were well above the target of .7.

Table 9. Grade 1 Standardized Factor Loadings for Initial and Revised Correlated-Trait Model

Factor	Indicator description	Initial model		Revised model	
		Estimate	(*SE*)	Estimate	(*SE*)
Counting					
Item 1					
		—	—	—	—
Item 2		.743	(.030)	.726	(.032)
Item 3					
		.894	(.032)	.906	(.035)
Word Problems					
Item 4		.501	(.041)	—	—
Item 5		.738	(.028)	.768	(.030)
Item 6		.516	(.036)	—	—
Item 7		.660	(.042)	.667	(.042)
Item 8		.616	(.032)	—	—
Item 9		.810	(.026)	.841	(.027)
Item 10		.722	(.030)	.757	(.030)
Computation					
Item 11		.687	(.026)	.668	(.027)
Item 12		.952	(.012)	.959	(.011)
Item 13		.690	(.031)	.678	(.032)
Item 14		.866	(.015)	.876	(.015)
Item 15		.663	(.032)	.655	(.032)
Item 16		.901	(.014)	.910	(.014)
Item 17		.793	(.021)	.801	(.021)
Item 18		.834	(.019)	.836	(.019)
Item 19		.564	(.030)	—	—
Item 20		.677	(.027)	.637	(.028)

Note. n= 1,226.

Table 10 presents the correlations among the factors for the grade 1 model. All interfactor correlations were statistically significant and moderate to large in size. No interfactor correlations were so large as to suggest colinearity. Figure 4 illustrates the correlated factor structure and standardized factor loadings for the revised grade 1 model.

Table 10. Grade 1 Factor Correlations (and Standard Errors) for the Revised Correlated-Trait Model

Factors	Counting	Word Problems	Computation
Counting	—		
Word Problems	.719 (.035)	—	
Computation	.556 (.035)	.578 (.026)	—

Note. n= 1,226
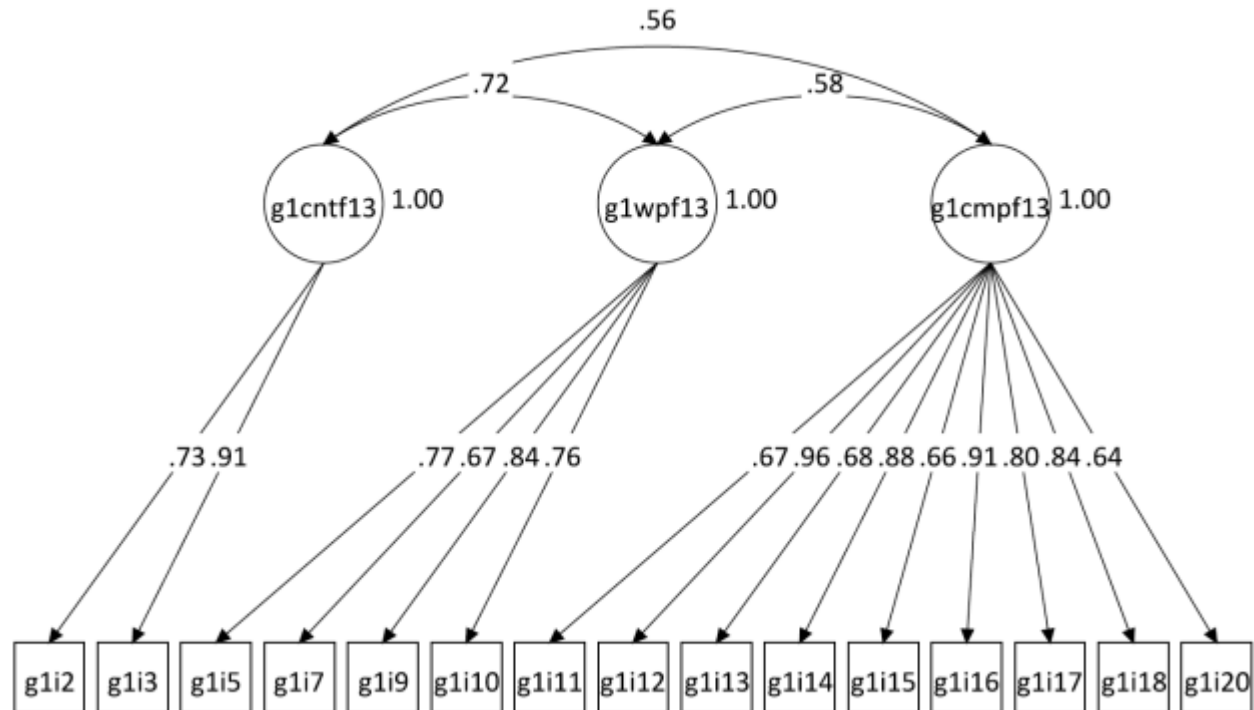
Figure 4. Grade 1 revised model—correlated-trait model diagram with standardized parameter estimates. Factor g1cntf13 is the grade 1 Counting factor for fall 2013. Factor g1wpf13 is the grade 1 Word Problems factor for fall 2013. Factor g1cmpf13 is the grade 1 Computation factor for fall 2013.

4.3.2. Grade 2 Correlated-Trait Model Evaluation

The initial grade 2 model contained all items that were administered. All items in the initial model had statistically significant unstandardized factor loading ($p < .001$). Seven items (4, 8, 10, 11, 12, 15, and 19) had standardized factor loadings that were near the factor loading minimum acceptable value of .5. Upon inspection of the standardized loadings for items 4 (.59), 8 (.54), 10 (.64),11 (.60), 12 (.52), 15 (.60), and 19 (.56) and their representation of the range of item difficulty, as well as consideration of their relative contribution toward the content validity of the scale, we determined that all of these items should be dropped for the revised model.

We then fit the data for the reduced set of grade 2 items to a revised correlated-trait structure and evaluated the factorial validity of the model on the basis of overall goodness of fit and interpretability, size, and statistical significance of the parameter estimates. The revised grade 2 correlated-trait model fit statistics indicated reasonable fit for the RMSEA statistic and close fit for the CFI and TLI statistics: $\chi^2(62) = 276.759$, $p < .001$; RMSEA = .055, 90% CI [.048, .062]; CFI = .962; and TLI = .952. All unstandardized factor loadings for the revised grade 2 model were statistically significant. Table 11 presents the standardized factor loadings for the initial and revised correlated-trait model. All standardized factor loadings for the revised grade 2 model were above the minimum acceptable value of .5, and most were well above the target of .7.

Table 11. Grade 2 Standardized Factor Loadings for Initial and Revised Correlated-Trait Model

Factor	Indicator description	Initial model Estimate	(*SE*)	Revised model Estimate	(*SE*)
Counting					
Item 1		.738	(.040)	.742	(.040)
Item 2		.789	(.030)	.798	(.030)
Item 3		.797	(.030)	.785	(.031)
Word Problems					
Item 4		.589	(.051)	—	—
Item 5		.788	(.025)	.811	(.027)
Item 6		.679	(.031)	.720	(.030)
Item 7		.811	(.027)	.842	(.029)
Item 8		.541	(.040)	—	—
Item 9		.648	(.031)	.671	(.032)
Item 10		.640	(.032)	—	—
Computation					
Item 11		.598	(.058)	—	—
Item 12		.524	(.044)	—	—
Item 13		.735	(.036)	.653	(.042)
Item 14		.710	(.031)	.756	(.030)
Item 15		.592	(.033)	—	—
Item 16		.753	(.024)	.798	(.024)
Item 17		.659	(.029)	.694	(.029)
Item 18		.761	(.025)	.801	(.026)
Item 19		.555	(.032)	—	—
Item 20		.654	(.029)	.677	(.029)

Note. n = 1,147.

Table 12 presents the correlations among the factors for the grade 2 model. All interfactor correlations were statistically significant and moderate to large in size. No interfactor correlations were so large as to suggest collinearity. Figure 5 illustrates the correlated factor structure and standardized factor loadings for the revised grade 2 model.

Table 12. Grade 2 Factor Correlations for the Revised Correlated-Trait Model

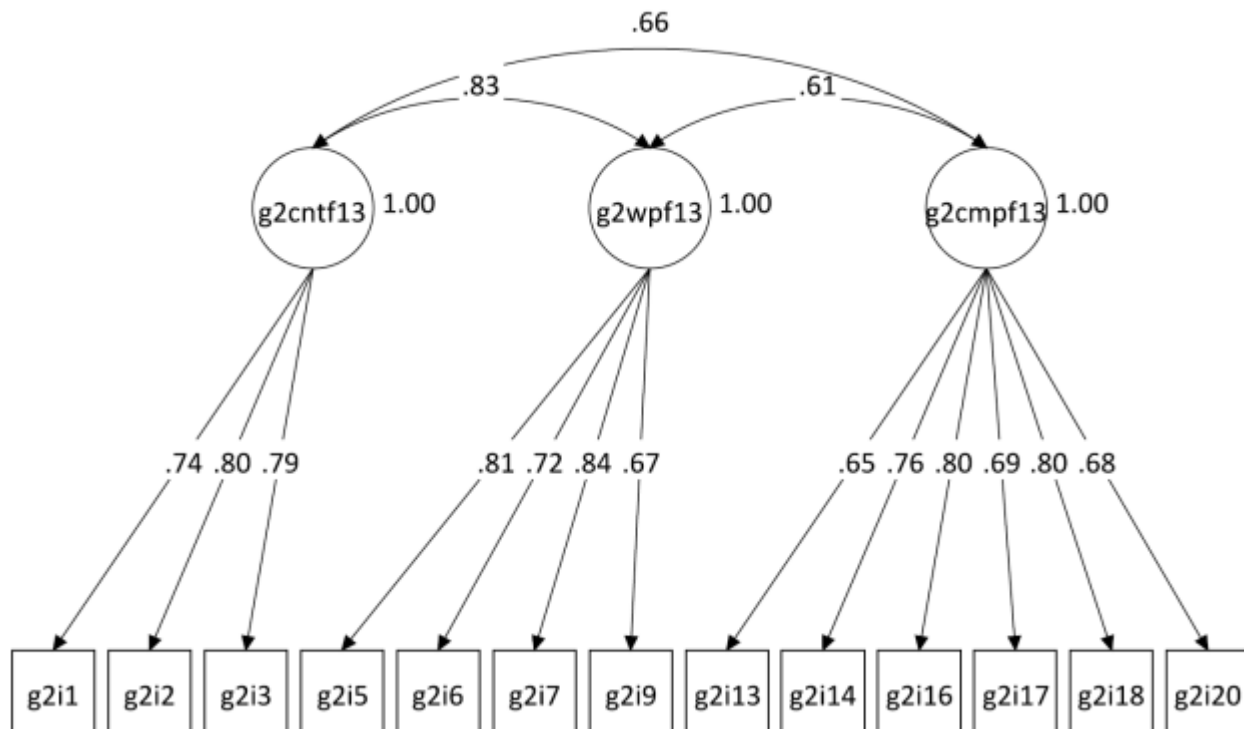| Factors	Counting	Word Problems	Computation
Counting	—		
Word Problems	.827 (.030)	—	
Computation	.663 (.037)	.606 (.033)	—

Note. n = 1,147

Figure 5. Grade 2 revised model—correlated-trait model diagram with standardized parameter estimates. Factor g2cntf13 is the grade 2 Counting factor for fall 2013. Factor g2wpf13 is the grade 2 Word Problems factor for fall 2013. Factor g2cmpf13 is the grade 2 Computation factor for fall 2013.

4.4. Higher-Order Model Evaluation

Higher-order factor models with three first-order factors are considered just identified. That is, the higher-order model and the correlated-trait model each use three parameters to specify the relationship between the first-order factors. Accordingly, which model fits the data better cannot be determined. Also, the fit statistics are identical for both structures, and the standardized factor loadings are nearly identical. Notwithstanding the indeterminacy of which model is better, the pragmatic advantage of using a higher-order factor structure to derive an overall score for the tests was compelling enough to justify its use for the final model.

4.4.1. Grade 1 Higher-order Model Evaluation

Table 13 presents the standardized factor loadings and factor residual variances for the grade 1 higher-order measurement model. Figure 6 illustrates the higher-order factor structure and standardized factor loadings for the final grade 1 model.

Table 13. Standardized Factor Loadings and Factor Residual Variances for the Grade 1 Higher-Order Measurement Model

Factor	Indicator description	Estimate	(*SE*)
	Lower-order factors		
Counting			
Item 1		—	—
Item 2		.726	(.032)
Item 3		.906	(.035)
Word Problems			
Item 4		—	—
Item 5		.768	(.030)
Item 6		—	—
Item 7		.667	(.042)
Item 8		—	—
Item 9		.841	(.027)
Item 10		.757	(.030)
Computation			
Item 11		.668	(.027)
Item 12		.959	(.011)
Item 13		.678	(.032)
Item 14		.876	(.015)
Item 15		.655	(.032)
Item 16		.910	(.014)
Item 17		.801	(.021)
Item 18		.836	(.019)
Item 19		—	—
Item 20		.637	(.028)
	Higher-order factor		
Math			
Counting	Counting latent variable	.832	(.038)
Word Problems	Word Problems latent variable	.864	(.034)
Computation	Computation latent variable	.668	(.028)
	Residual variance		
Counting		.308	(.063)
Word Problems		.253	(.058)
Computation		.553	(.037)

Note. n = 1,226.
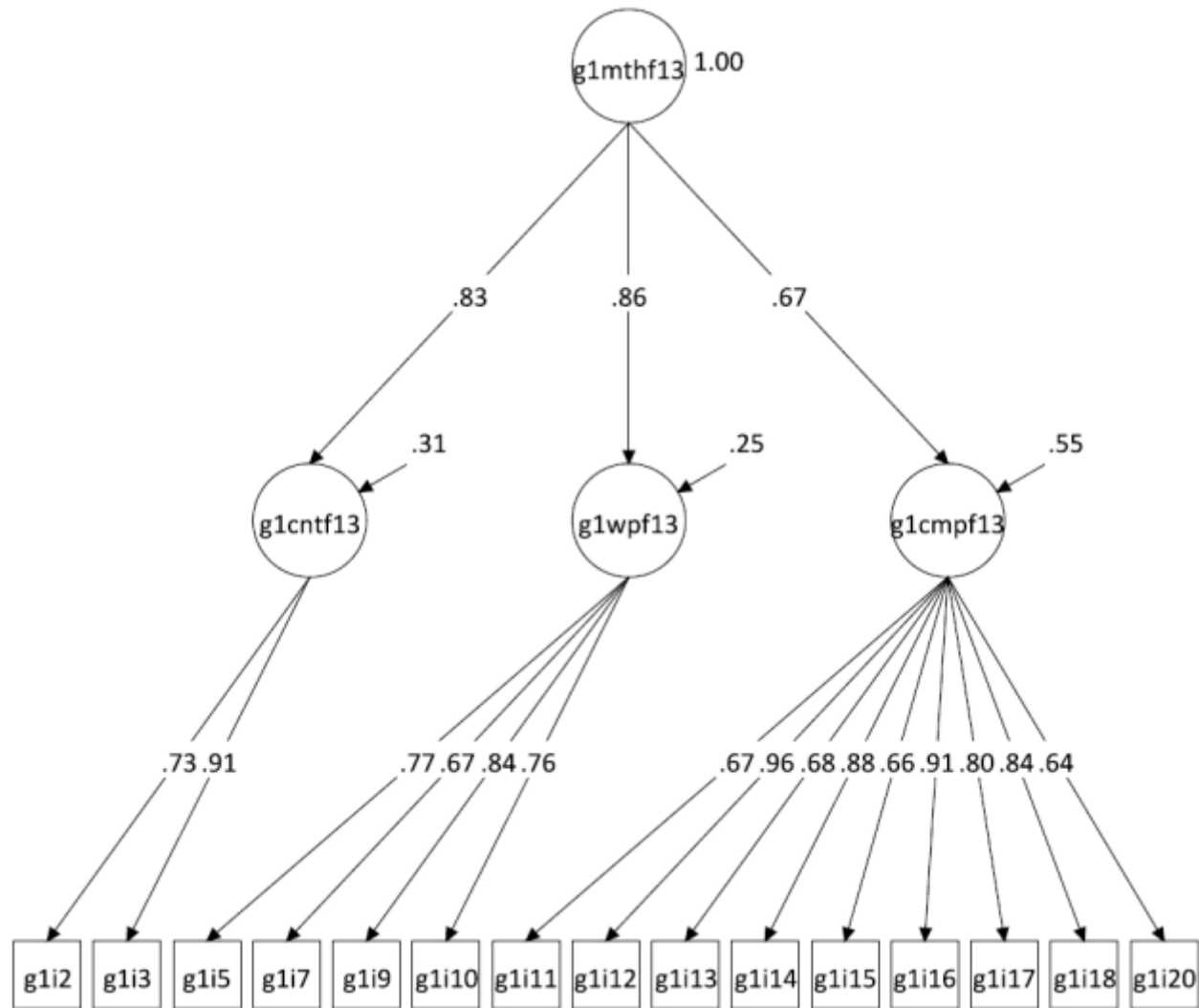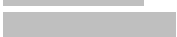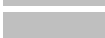
Figure 6. Grade 1 final model—higher-order factor diagram with standardized parameter estimates.

4.4.2. Grade 2 Higher-order Model Evaluation

Table 14 presents the standardized factor loadings and factor residual variances for the grade 2 higher-order measurement model. Figure 7 illustrates the higher-order factor structure and standardized factor loadings for the final grade 2 model.

Table 14. Standardized Factor Loadings and Factor Residual Variances for the Grade 2 Higher-Order Measurement Model

Factor	Indicator description	Estimate	(*SE*)
	Lower-order factors		
Counting			
Item 1		.742	(.040)
Item 2		.798	(.030)
Item 3		.785	(.031)
Word Problems			
Item 4		—	—
Item 5		.811	(.027)
Item 6		.720	(.030)
Item 7		.842	(.029)
Item 8		—	—
Item 9		.671	(.032)
Item 10			
Computation			
Item 11		—	—
Item 12		—	—
Item 13		.653	(.042)
Item 14		.756	(.030)
Item 15		—	—
Item 16		.798	(.024)
Item 17		.694	(.029)
Item 18		.801	(.026)
Item 19		—	—
Item 20		.677	(.029)
	Higher-order factor		
Math			
Counting	Counting latent variable	.952	(.033)
Word Problems	Word Problems latent variable	.869	(.029)
Computation	Computation latent variable	.697	(.032)
	Residual variance		
Counting		.095	(.063)
Word Problems		.244	(.050)
Computation		.514	(.044)

Note. n = 1,147.
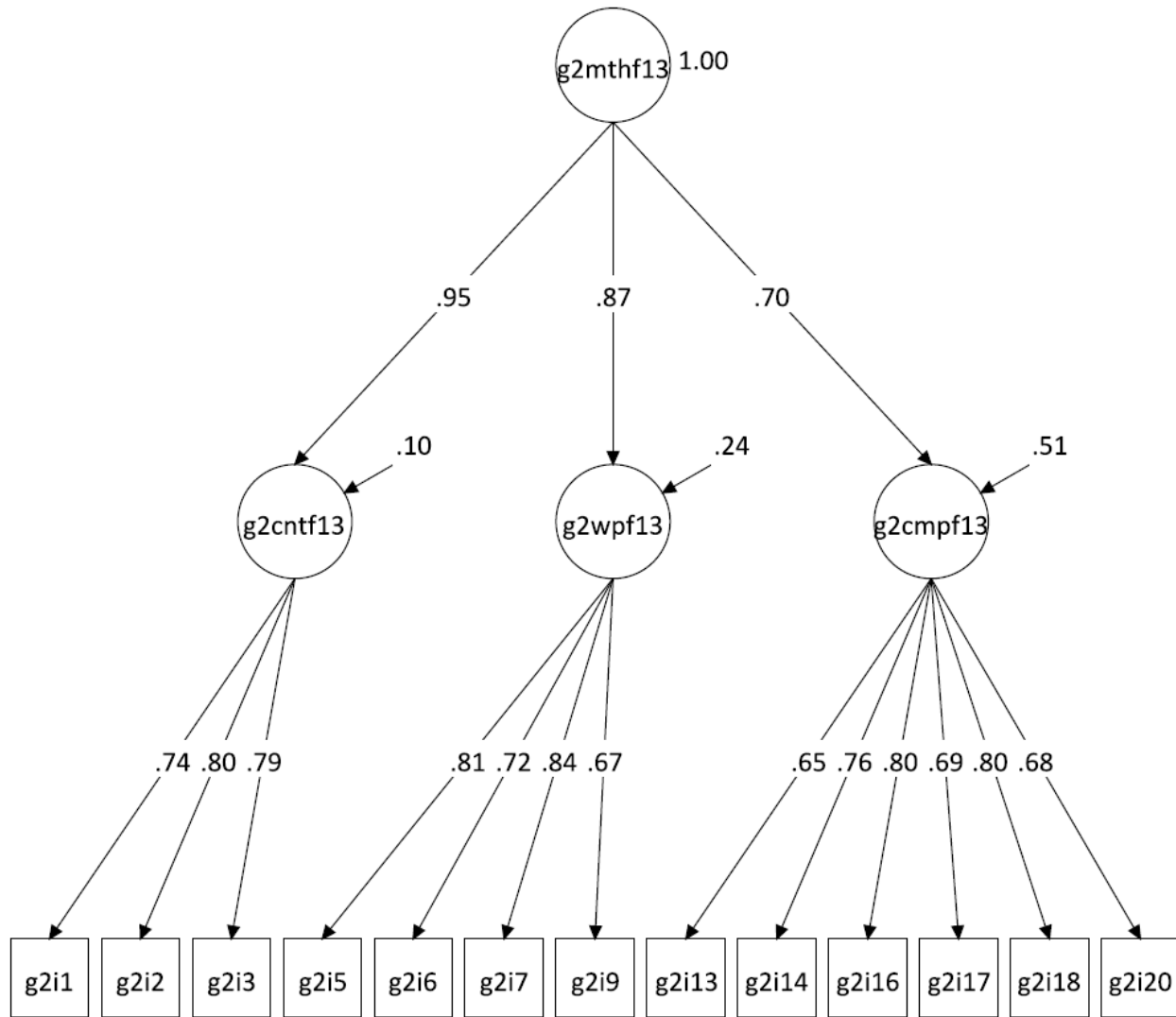
Figure 7. Grade 2 final model—higher-order factor diagram with standardized parameter estimates.

4.5. Scale Reliability Evaluation

4.5.1. Grade 1 Scale Reliabilities

The scale reliabilities for the grade 1 test suggested acceptable reliability for all scales. The grade 1 higher-order Math factor composite reliability estimate was evaluated by means of Equation 3, where the numerator is the squared sum of the unstandardized second-order factor loadings and the denominator is the squared sum of the unstandardized second-order factor loadings plus the sum of the first-order factor residual variances.

$$\frac{(0.754 + 0.654 + 426)^2}{(0.754 + 0.654 + 426)^2 + (0.253 + 0.145 + 0.224)} = .844 \tag{3}$$

The present sample indicated a composite reliability of .84 for the grade 1 higher-order Math factor, which exceeds the target reliability of .8.

Table 15 presents the α, β, and ω_h ordinal reliability coefficients for the reduced set of items by subscale and for the total scale. The α estimates for the Word Problems and Computation scales exceeded the target of .8. The estimated α reliability of the Counting scale was .79. Comparison between the αs and βs revealed a range of discrepancies, some moderate (e.g., for the Word Problems scale, where α = .84 and β = .78) and others large (e.g., for the Computation scale, where α = .91 and β = .60). The magnitudes of discrepancies indicate heterogeneity among the factor loadings, challenging the assumption of essential tau equivalence. Comparison between the α and ω_h coefficients revealed discrepancies to be moderate (.05) for the Word Problems scale and large for the Computation scale (.32) and total Math scale (.22). (An ω_h coefficient could not be computed for the Counting scale because the scale included only two items.) For all estimates, α exceeded ω_h, with the α to ω_h discrepancies indicating the presence of multidimensionality within the scales. The Word Problems and total Math scales' ω_h met or exceeded the conventional minimum value of .7, suggesting composite scores can be interpreted as reflecting a single common source of variance in spite of evidence of some within-scale multidimensionality (Gustafsson & Aberg-Bengtsson, 2010). The ω_h for the Computation scale did not, however, exceed the conventional minimum threshold, indicating the presence of substantial within-scale multidimensionality for that scales.

Table 15. Grade 1 Scale Reliability Estimates

Scale	Number of items	Reliability		
		α	β	ω_h
Counting	2	.79	.79	—
Word Problems	4	.84	.78	.79
Computation	9	.91	.60	.59
Math	15	.92	.77	.70

Note. n = 1,226. α, β, and ω_h are ordinal forms of Cronbach's α, Revelle's β, and McDonald's ω_h hierarchical, respectively.

Inspection of the 2-pl UIRT TIC in Figure 8, reveals that the information curve for the grade 1 test exceeded 2.33 (reliability of .7) for the ability range of approximately −1.4 through 1.9. Given the sample descriptives (M = 0.00, SD = 0.92, Min = −2.00, and Max = 2.02), this result suggests acceptable reliability of the scale for approximately 92% of the sample and nearly the full range of observed abilities. The information curve exceeded 4 (reliability of .8) for the ability range of approximately −0.6 through 1.5, indicating that target reliability of the scale was achieved for approximately 70% of the sample.[1] The information curve exceeds 4 (reliability of .8) for the ability range of approximately −1.8 through 0.5, indicating target reliability of the scale was achieved for approximately 69% of the sample.

---

[1]Areas under normal distribution calculated with the online normal distribution calculator found at http://onlinestatbook.com/2/calculators/normal_dist.html

Figure 8. Grade 1 2-pl UIRT total information curve and participant descriptives for the reduced set of items modeled as a single factor.

Figure 9 presents the overall distribution of number of items answered correctly in grade 1 for the reduced set of items. Similar figures for each subscale are provided in Appendix E.

Figure 9. Distribution of the number of items individual students in the grade 1 sample answered correctly on the reduced set of items.

4.5.2. Grade 2 Scale Reliabilities

The scale reliabilities for the grade 2 test suggested acceptable reliability for all scales. The grade 2 higher-order (i.e., Math) factor composite reliability estimate was calculated from Equation 4, where the numerator is the squared sum of the unstandardized second-order factor loadings and the denominator is the squared sum of the unstandardized second-order factor loadings plus the sum of the first-order factor residual variances.

$$\frac{(0.747 + 0.583 + 472)^2}{(0.747 + 0.583 + 472)^2 + (0.058 + 0.110 + 0.236)} = .889 \tag{4}$$

We calculated a composite reliability for the grade 2 higher-order Math factor of .88, which exceeds the target reliability of .8.

Table 16 relays the α, β, and ω_h ordinal reliability coefficients for the reduced set of items by subscale and for the total scale. All α estimates for all subscales exceeded or met the target of .8. As with the grade 1 test, comparison between the αs and βs revealed a range of discrepancies (range .00 to .14), challenging the assumption of essential tau equivalence where the discrepancy was sizable. Comparison between the α and ω_h coefficients also revealed a range of discrepancies (range .00 to .18). Where α exceeded ω_h (i.e., Word Problems, Computation, and Math), the α to ω_h discrepancies indicate the presence of multidimensionality within the scales. Where ω_h was equal to α (i.e., Counting), it means there was variability in the general factor loadings but group factor loadings were relatively small, indicating that lumpiness in the scale is not attributable to multidimensionality. In every case, ω_h exceeded the conventional minimum value of .7. As demonstrated by Gustafsson and Aberg-Bengtsson (2010), high values of ω_h indicate that composite scores can be interpreted as reflecting a single common source of variance in spite of evidence of some within-scale multidimensionality.

Table 16. Grade 2 Scale Reliability Estimates

Scale	Number of items	Reliability		
		α	β	ωh
Counting	3	.82	.73	.82
Word Problems	4	.85	.85	.83
Computation	6	.86	.80	.74
Math	13	.91	.77	.73

Note. n = 1,147. α, β, and ωh are ordinal forms of Cronbach's α, Revelle's β, and McDonald's ωh, respectively.

Inspection of the 2-pl UIRT TIC in Figure 10, reveals the information curve for the grade 2 test to exceed 2.33 (reliability of .7) for the ability range of approximately –2.4 through 1.0. Given the sample descriptives (M = 0.00, SD = 0.89, Min = -2.32, and Max = 1.39), reliability of the scale is therefore acceptable for over 87% of the sample and nearly the full range of observed abilities. The information curve exceeds 4 (reliability of .8) for the ability range of approximately –1.8 through 0.5, indicating that target reliability of the scale was achieved for approximately 69% of the sample.
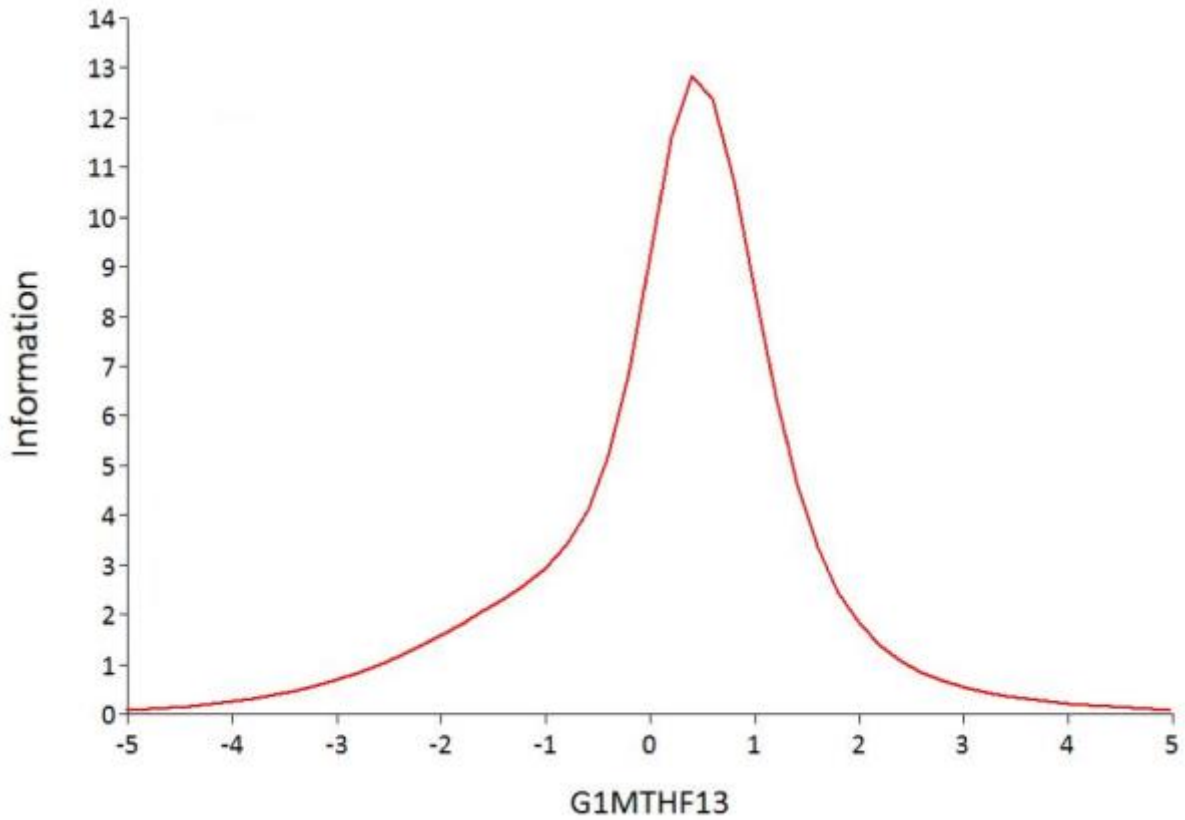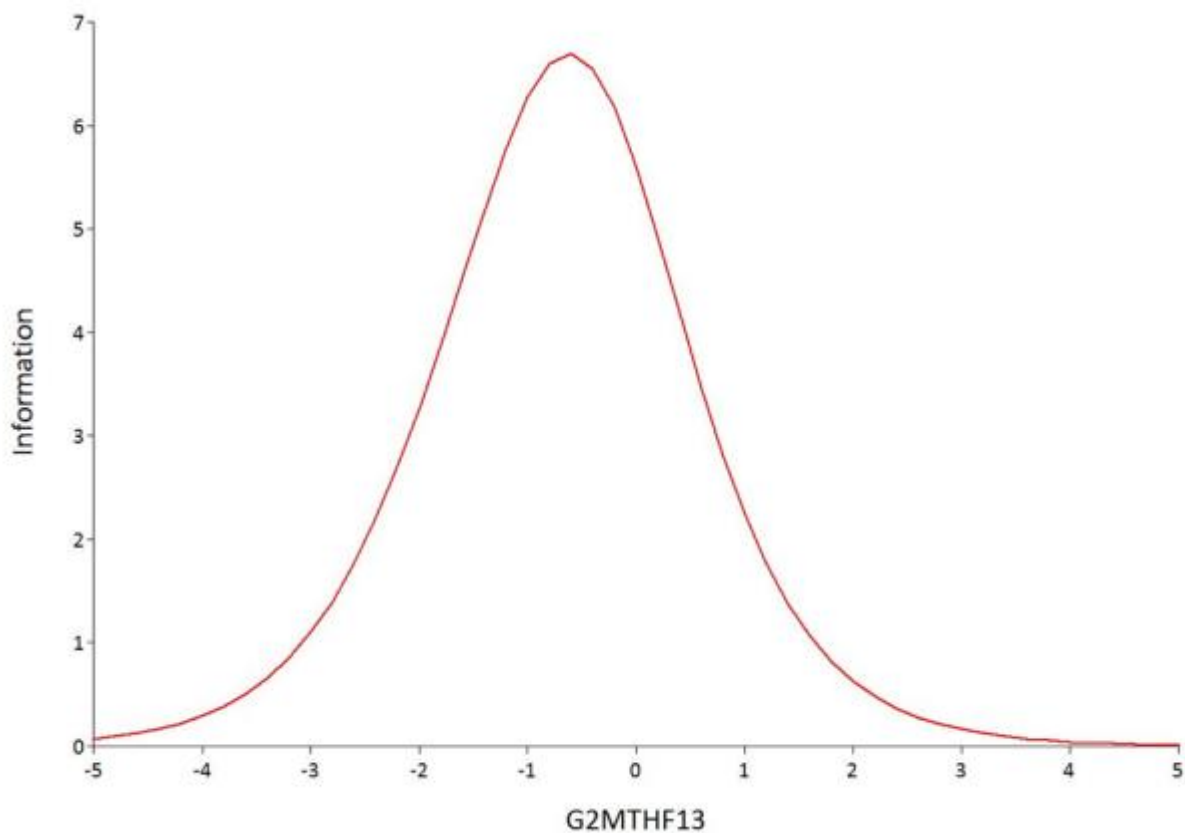

Figure 10. Grade 2 2-pl UIRT total information curve and participant descriptives for the reduced set of items modeled as a single factor.

Figure 11 presents the overall distribution of number of items answered correctly in grade 2 for the reduced set of items. Similar figures for each subscale are provided in Appendix G.
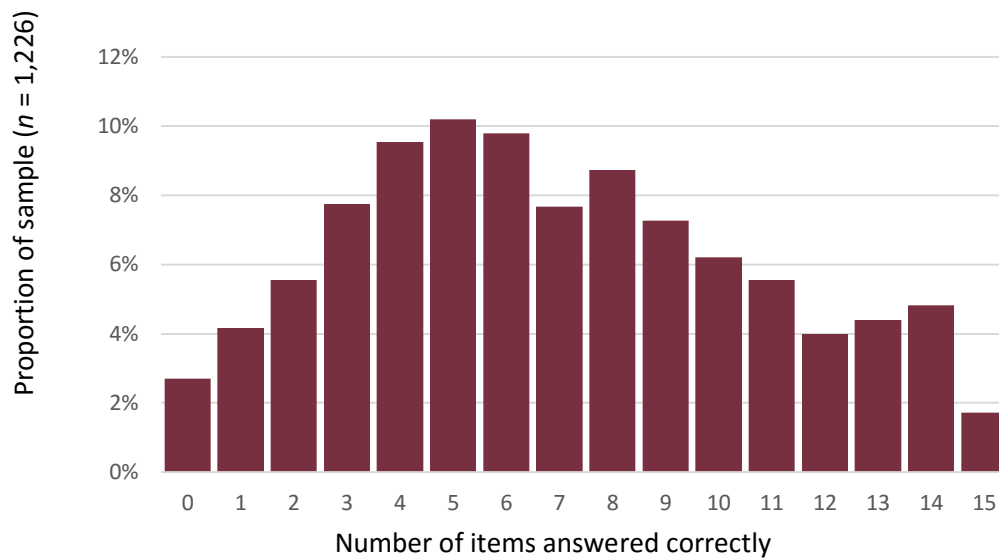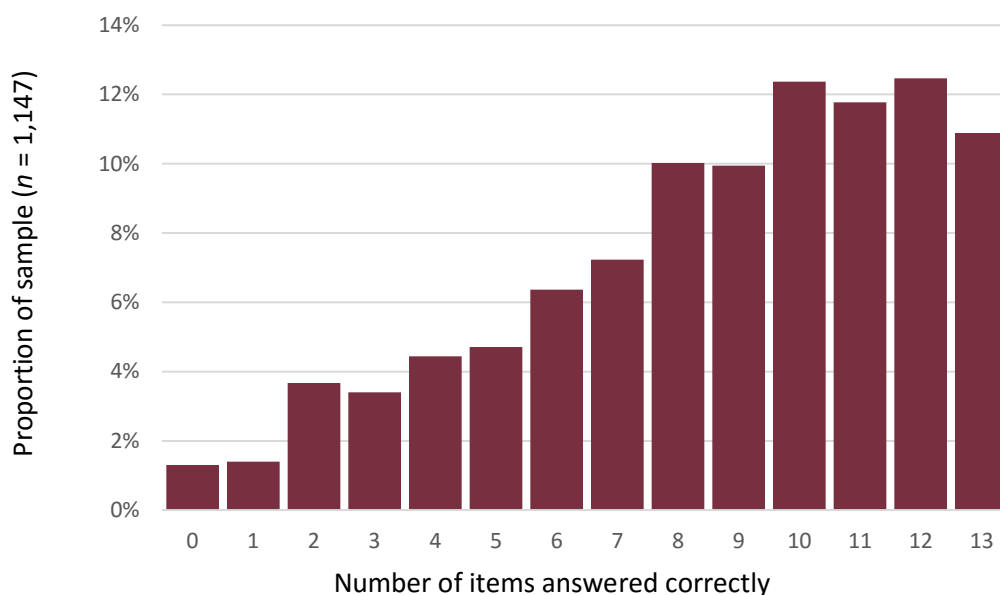

Figure 11. Distribution of the number of items individual students in the grade 2 sample answered correctly on the complete reduced set of items.

4.6. Validity Evaluation

4.6.1. Concurrent Validity Evaluation

All correlation coefficients were moderate in size (*r* range .32 to .69) and statistically significant at *p* < .001. With a correlation coefficient of *r* = .69, only the correlation between the grade 1 test total Math factor score and the grade 1 DEA overall scale score approached the .7 threshold for scale concordance. The correlation between the grade 2 test total Math factor score and the grade 2 DEA overall scale score was *r* = .61. Notwithstanding attenuation of correlations due to scale reliability, the statistically significant, moderately-sized correlation coefficients provide some, albeit modest, evidence of concurrent validity. Table 17 presents the coefficients for the correlations between the student test and the DEA for each grade.

Table 17. Correlations among Test Scales and the DEA for each Grade

DEA overall score and subdomains	Researcher-developed student test subdomains			
	Counting	Word Problems	Computation	Math
Grade 1				
Overall scale score	.654	.677	.530	.691
Operations	.434	.436	.348	.451
Base Ten	.454	.510	.377	.503
Measurement and Data	.606	.588	.489	.620
Grade 2				
Overall scale score	.602	.601	.449	.610
Operations	.498	.513	.322	.504
Base Ten	.481	.480	.400	.491
Measurement and Data	.508	.500	.395	.514

Note. Grade 1 DEA *n* = 320. Grade 2 DEA *n* = 351. All correlations were statistically significant at *p* < .001.

4.6.2. Predictive Validity Evaluation

We used regression analyses to explore the extent to which the EMSA Math factor predicted performance on each of the two ITBS tests (i.e., Math Problems, Math Computation) at each grade level. Regression results suggested that the test total Math score was a moderate to strong predictor of the ITBS Math Problems test, where an $R^2_{Adjusted}$ of .41 was found for the grade 1 control group and an $R^2_{Adjusted}$ of .49 was found for the grade 2 control group. The test total Math score provided only modest predictive power with the ITBS Math Computation test, where an $R^2_{Adjusted}$ of .23 was found for the grade 1 control group and an $R^2_{Adjusted}$ of .30 was found for the grade 2 control group. All models were statistically significant at *p* < .001. Table 18 presents the results for the single linear regressions of the ITBS Math Problems and Math Computation tests on the test total Math scale when they were applied to the grade 1 and grade 2 control group.

Table 18. Results for Single Linear Regressions of Standard Scores on the Iowa Test of Basic Skills (ITBS) Math Problems and Math Computation Tests on the Math Factor Scores for the Grade 1 and Grade 2 Control Group

Criterion	*df*		*F*			
	Regression	Residual	Statistic	*p*	β	$R^2_{Adjusted}$
Grade 1 control group						
ITBS Math Problems	1	489	347.623	< .001	.645	.414
ITBS Math Computation	1	489	143.808	< .001	.477	.226
Grade 2 control group						
ITBS Math Problems	1	468	456.712	< .001	.703	.494
ITBS Math Computation	1	468	194.052	< .001	.547	.298

Note. Grade 1 ITBS Math Problems *n* = 491. Grade 2 ITBS Math Problems *n* = 470.

5.1 Discussion and Conclusions

The intended use of the fall 2013 EMSA tests was as a baseline test of student achievement to be used as a covariate in a randomized controlled trial of a teacher professional-development intervention and as a pretest student achievement measure to test for baseline equivalence of the schools assigned to the treatment and control conditions. The development and analysis of the fall 2013 EMSA tests are consistent with general recommendations for test development and test validation for the intended purposes of the Fall 2013 EMSA. To be used for other purposes, such as to distinguish among levels of individual student achievement, the test would require further development and validation.

The field test of the fall 2013 EMSA tests involved a diverse sample of several thousand grade 1 and 2 students in fall 2013. The tests were administered at the beginning of the school year by classroom teachers in most cases. The test scores were not known by the schools or used for any kind of school- or teacher-accountability purpose. Our sample does not reveal how changes in these testing conditions might affect the data. Further validation efforts would be necessary if the test were administered under different conditions or used for different purposes.

5.1. Validation

5.1.1. Substantive Validation

The analysis of content in the CCSS-M and the CGI professional-development program provided guidance for the content of the fall 2013 EMSA tests. Administration procedures were consistent with typical classroom assessment in mathematics, including that of standardized tests such as the ITBS. External review of items and scoring criteria provided further support for the substantive phase of construct validation.

5.1.2. Structural Validation

The structural phase of validation was fairly extensive in the field test of the fall 2013 EMSA tests. Initial screening provided a calibration phase to adjust the difficulty and discrimination of items to the target population. The data were fit to both a correlated-traits and a second-order factor analysis model. To generate overall test scores, three first-order factors (Counting, Word Problems, Computation) were regressed onto a single second-order factor (Math). The second-order Math factor score is intended to serve as the overall achievement score on the test. Goodness-of-fit statistics varied, though they generally indicated that the specified measurement models provided a reasonable fit to the data. All unstandardized factor loadings for both models were statistically significant.

The reliability estimates for both of the test scales met standards for educational research. Little discrepancy was apparent among these various reliability estimates (e.g., ordinal forms of Revelle's β and McDonald's ω_h coefficients), but the McDonald's ω_h for the higher-level Math factor can be interpreted to indicate potential multidimensionality in the scale.

5.1.3. External Validation

Moderate correlations between the fall 2013 EMSA test scores and the fall DEA (2010) test scores were observed. Notwithstanding attenuation of correlations due to scale reliability, the statistically significant, moderately-sized correlation coefficients provide some, albeit modest, evidence of concurrent validity.

For its intended use, knowing the proportion of variance in posttest scores is explained by pretest scores can be particularly useful to researchers and evaluators in the power analysis phase of research design. The regression analyses suggest the test to be an appropriate covariate in analyses that use the ITBS tests as outcomes, where the results suggest the test is particularly well suited in analyses with the ITBS Math Problems test.

5.2. Improving the Test

One way to improve the reliability and alignment of the test with student abilities may be to replace some of the items on the grade 2 test that were not included in the final scale with items that have higher difficulty levels. Conversely, the items on the test form that were not included in the final scale for the grad 1 test might be replaced by items with slightly lower difficulty levels.

An area for improvement and further development could be to design the tests so that they can be linked vertically across grade levels (using a common set of anchor items in each of the three sections of the test) to enable the grade 1 and 2 scores to be generated on a common scale. Vertical scaling would permit pooling of data across grade levels, which might increase statistical power for a given sample involving students at multiple grade levels.

Test specifications indicated that images from openclipart.com would be used for page-numbering (rather than numerals, which could potentially confuse or mislead the young children taking the test). In several cases, the subject of a word problem (e.g., balloons, books) was used as the image on the same page as the word problem. In retrospect, this decision may have created confusion, especially when the number of balloons in the image matched a quantity in the problem. In the future, this page-numbering technique will continue to be used, but we will not use an image that corresponds directly to the objects in the word problem.

5.3. Summary and Conclusions

The development process and results of the field test of the fall 2013 EMSA provide evidence of substantive, structural, and external validity of the fall 2014 EMSA tests (Flake, Pek, & Hehman, 2017). The fall 2013 EMSA tests were field-tested with more than 1,200 grade 1 students and more than 1,100 grade 2 students. Reliability estimates suggest that the test may be adequately reliable for its intended purpose. The results of the field-test indicate that the fall 2013 EMSA tests are well suited for their intended purpose.

References

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258.

Camilli, G. (1994). Origin of the scaling constant d = 1.7 in item response theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293–295.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction.* Portsmouth, NH: Heinemann.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*(4), 385–531.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462-494.

DEA (Discovery Education Assessment) (2010). *Discovery Education's Common Core Mathematics grade 1 and grade 2 interim benchmark assessment*. Silver Spring, MD: Discovery Education.

Dunbar, S. B., Hoover, H. D., Frisbie, D. A., Ordman, V. L., Oberley, K. R., Naylor, R. J., & Bray, G. B. (2008). *Iowa Test of Basic Skills,® Form C, Level 7*. Rolling Meadows, IL: Riverside Publishing.

Embretson, S.E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27*(4), 458–477.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 1–9.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3). Available online: http://pareonline.net/getvn.asp?v=17&n=3

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91.

Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97–121). Washington, DC: American Psychological Association.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55, doi: 10.1080/10705519909540118.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education, 38*(3), 258–288.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

NGACBP (National Governors Association Center for Best Practices) and CCSSO (Council of Chief State School Officers) (2010). *Common Core State Standards for Mathematics*. Washington, DC: Authors.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Nye, C. D. & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3), 548–570.

R Development Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Reise, S. P., Horan, W. P., & Blanchard, J. J. (2011). The challenges of fitting an item response theory model to the Social Anhedonia Scale. *Journal of Personality Assessment, 93*(3), 213–224.

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*(1), 57–74.

Revelle, W. (2016). *psych: Procedures for personality and psychological research* (Version 1.6.6). Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Schoen, R. C., Champagne, Z. M., Whitacre, I., & McCrackin, S. (Manuscript under review). Comparing the frequency and variation of additive word problems in U.S. first-grade textbooks in the 1980s and the Common Core era.

Schoen, R. C., LaVenia, M., Champagne, Z. M., & Farina, K. (2016). Mathematics performance and cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2014. (Research Report No. 2016–01.) Tallahassee, FL: Learning Systems Institute. doi:10.1725/fsu.1493238156

Stigler, J. W., Fuson, K. C., Ham, M., & Kim, M. S. (1986). An analysis of addition and subtraction word problems in American and Soviet elementary mathematics textbooks. *Cognition and Instruction, 3*(3), 153–171.

Streiner, D. L. (2003) Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.

Appendix A—First Grade Test

**Student Mathematics Assessment
First Grade**

District:
School:
Teacher: Grade:
Student:

Sample <u>fill in the bubble</u> multiple-choice

What grade are you in?

K   1   2   3   4

Sample <u>write in the box</u>

Write the number four in the box:
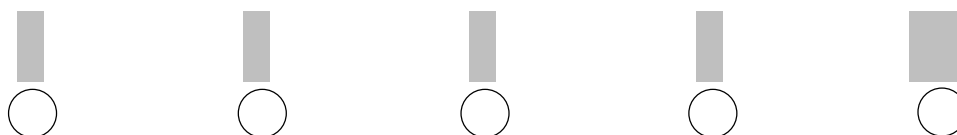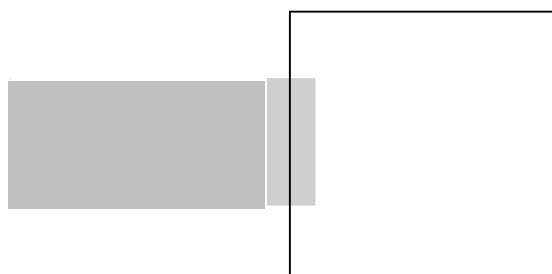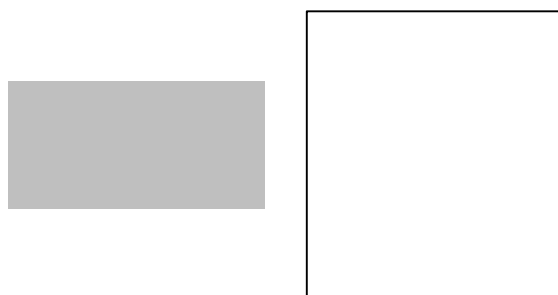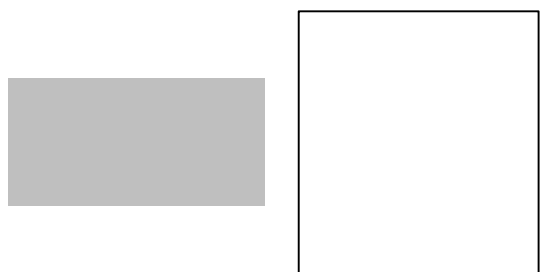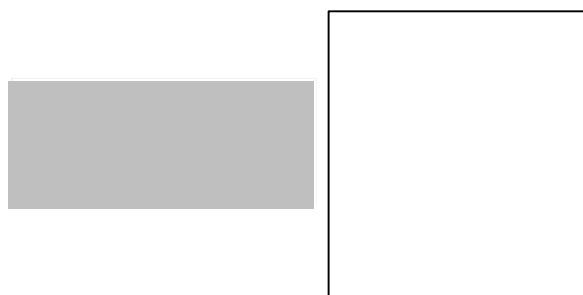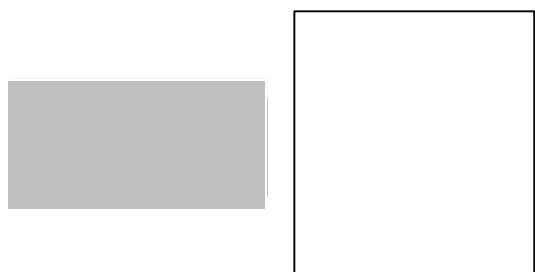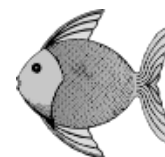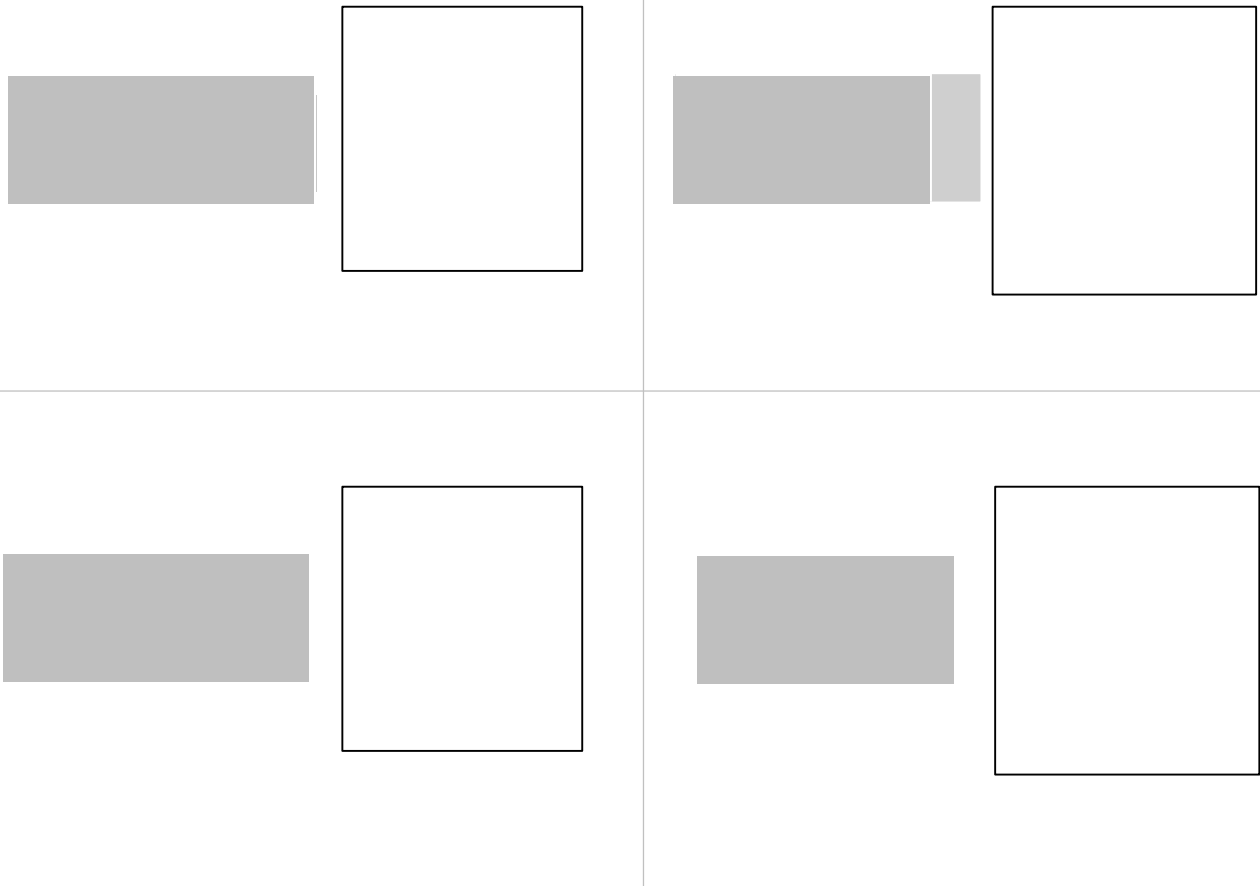
Appendix B—Second Grade Test

**Student Mathematics
Assessment Second Grade**

District:

School:

Teacher: Grade:

Student:

Sample <u>fill in the bubble</u> multiple-choice
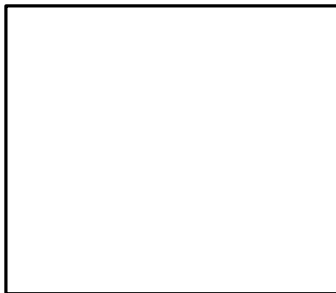
What grade are you in?

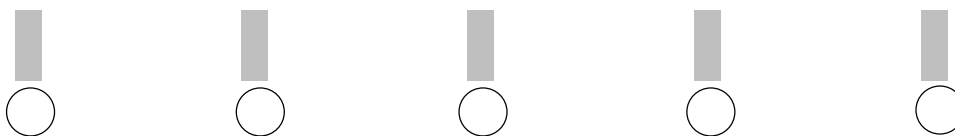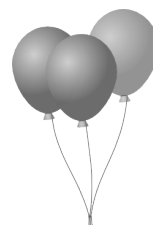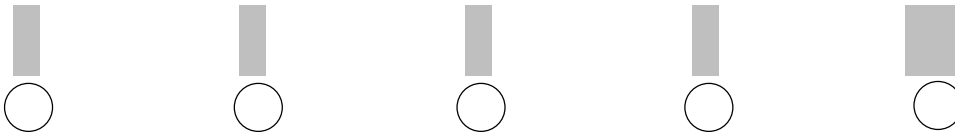| K	1	2	3	4
○	○	⬤	○	○

Sample <u>write in the box</u>

Write the number four in the box: 4
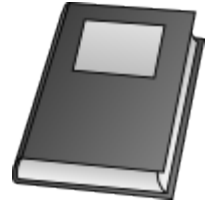
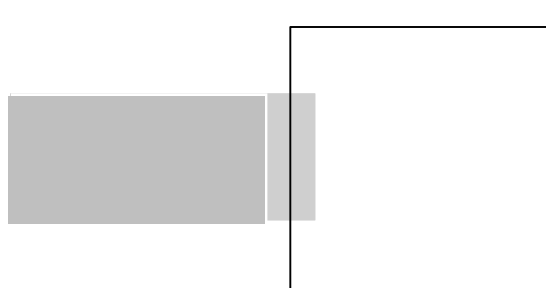# Appendix C—First Grade Administration Guide

Primary Grades Math Study:

Pre-test Guidelines, Administration Instructions, and Student Information Sheet

Grade 1

2013- 2014

Table of Contents

Pre-test Guidelines

Overview

The Primary Grades Math Study pre-test (hereafter, pre-test) provides four sections of assessments: Counting, Writing Numerals, Solving Word Problems and Addition/Subtraction Number Facts (First Grade).

The following guidelines provide information on the protocol for administering the pre-test. Throughout this document a second-person voice is used with the intended reader being the classroom teacher. It is assumed that the classroom teacher will administer the pre-test; however, it is permissible for other school personnel (such as a paraprofessional or even a substitute teacher) to administer the pre-test, providing they follow the pre-test protocol as detailed below.

Pre-test Testing Window

Local Education Agency	Testing Window
District A	August 19 – August 30, 2013
District B	August 12 – August 23, 2013

Please identify your locale in the table below for the applicable testing window.

Materials

The following materials are <u>required</u> for testing:

- Primary Grades Math Study Pre-test Guidelines and Administration Instructions (provided)
- A test booklet for each student (provided)
- At least one sharpened pencil for each student

The following materials are <u>encouraged</u> for testing:

- Counters and/or linking cubes for each student

Test Booklets

Test booklets are consumable and students mark their answers directly in the test booklets. Should you need additional testing materials, please contact Kristopher Childs (<u>kristopher.childs@ucf.edu</u>). Remember that these materials are to remain at the school site until the testing window has ended. The materials must be stored in a secure, access-restricted location at all times.

Students to be Tested

The pre-test for the Primary Grades Math Study will be administered to students who have returned signed consent forms indicating parental consent to participate in the study. On the pre-test student information sheet (p. 13 of this document), please list only those students for whom you have signed consent and provide their information in the table as requested. Only pre-tests completed by these students are to be relayed to project personnel.

At your discretion, the pre-test may also be administered to students who have not returned consent forms, with the understanding that students may return consent forms after the pre-test has been administered. In such a case, please retain possession of those students' pre-tests until such time that it is certain that parental consent is not granted.

Also at your school's discretion, the pre-test may be administered to students whose parents have declined consent to participate, so long as you do not relay their materials or data to project personnel. That is, you are free to use this test like you would any other test to assess your students' mathematics ability. Accordingly, it may make most sense to administer it to your whole class, irrespective of their status with the study—understanding that you will only relay materials for students who have signed parental consent.

Preparing for Testing

The first page of each test booklet has the following box for student information:

District:

School:

Teacher: Grade:

Student:

Prior to the testing session, the classroom teacher must enter this information (district name, school name, teacher name, student name, and student grade level) on each test booklet for each student to be tested. (Please do not leave it for students to enter this information.)

The pre-test for Primary Grades Math Study may be administered to students on either an individual or group basis. Please adhere to the following guidelines:

1. Ensure all students have testing materials (i.e., test booklet and a sharpened pencil).
2. Ensure that students and pre-labeled test booklets are properly paired (i.e., each student receives the test booklet that has his or her name written on it).
3. Provide students with a comfortable testing environment.
4. Testing administrators should adhere to the pre-test guidelines and administration Instructions.
5. No talking or communication between students is permitted during testing.
6. Students are permitted to use mathematics manipulatives during the pre-test.

Manipulatives

If students would ordinarily be permitted to use manipulatives in your classroom to solve math problems, then they should also be permitted for the pre-test.

Administering the Test

The testing conditions for the pre-test should be consistent with the testing conditions for other student assessments administered in the classroom. For example, students should space out the desks or use student "privacy folders" if that is what they would usually do.

Avoid reading problems or answering student questions in a way that may offer clues to the correct answer. Student responses should reflect their current math knowledge. Thus, it is important that effort is taken to ensure that the test questions are clearly presented and that students understand how they are to mark their answer; however, great care should be taken to not lead students to the correct answer. To ensure that the students' test responses are valid, it is important that appropriate procedures are followed when administering the pre-test. These procedures include:

- Administration of the appropriate test level (Grade 1 pre-test for Grade 1 students, etc.)

- Adherence to the pre-test guidelines and administration instructions in order to provide a standardized testing protocol across classrooms
- Maintenance of test security

Accommodations

Students with special academic plans (e.g., IEP, 504, ELL) may receive whatever accommodations are specified in their plans, at the teacher's discretion.

Testing in the Primary Grades

It is understood that children at this age level vary in their familiarity with whole-group testing procedures. The following recommendations are provided to facilitate a smooth testing procedure and minimize student frustration:

- Ensure students understand the testing instructions.
- Monitor students to ensure they are completing the correct question.
- Provide students with sufficient time to answer the questions.

When Students Get Stuck on a Problem

The following are suggested solutions for when students appear stuck and do not mark [or write] an answer for a given problem. Start with the first suggestion and only go on to subsequent suggestions if the prior ones did not resolve the student's delay in marking an answer:

1) Ask the student(s), "Would you like me to read the problem again?" Re-read the problem and accompanying directions if requested to do so.
2) Ask the student(s), "Do you have a question about how to mark your answer?" If the student answers in the affirmative, reiterate the directions from the first page on how to fill in the bubble or write in the box; whichever is appropriate for the given problem.
3) State, "I'm going to wait for another minute before going on to the next problem. Please look at the problem and mark [or write] what you think is a correct answer to that problem."
4) After waiting another minute, restate the direction to mark the answer for that given problem (for example, "Fill in the bubble that goes with your answer"), then read from the top of the script box for the next problem.
5) Tell the student it is okay if he or she skips that problem for now. He or she can come back to it after finishing the rest of the problems – if there is time.

Testing Time Allocation

Administration of the pre-test should take approximately 45 minutes. This is not a timed test, and students should be allowed adequate time to answer the test questions.

Submitting the Pre-test Materials

Upon conclusion of testing, separate out the test booklets for those students who have returned signed parental consent for participation in the study and repack them in the original packaging. Please be sure to include the pre-test guidelines, administration instructions, and completed student information sheet in the package. All unused test booklets should be repacked for return to project personnel. A Primary Grades Math Study representative will coordinate with your school to set a date to retrieve the testing materials from you. The target period of pickup will be the week of September 9.

Remaining test booklets will be for either those students of whom their parents have declined consent or have yet to return consent forms at all. Please retain the test booklets from this latter group of students (i.e., have not returned the consent form), in the event that they do bring back signed parental consent over the coming days or weeks. At that time, you will transfer their test booklet to a Primary Grades Math Study representative. If you have questions about this process, contact kristopher.childs@ucf.edu. To maintain the security of the test, please dispose of the test booklets for students whose parents have declined consent.

Pre-test Administration Instructions – Grade 1

[The boxes contain the script that you will read to the student.]

> Your class is about to take a short math assessment. You will need a pencil.

Verify that all students have a pencil.

> I will now pass out the assessments. The assessments are already labeled with your names. When you receive the assessment, keep it face up, and do not turn any pages; we will all begin at the same time after I go over the instructions.

Ensure that students and pre-labeled test booklets are properly paired (i.e., each student receives the test booklet that has his or her name written on it).

State the following box only if manipulatives are being used during pre-test administration.

> For the math assessment it may help you to use manipulatives. I have placed manipulatives [indicate location of manipulatives]. The manipulatives can be used at any time during testing.

> The first page of the assessment gives the instructions and provides samples of how you will mark your answers.
>
> For some problems you will fill in the bubble beneath (below) the answer choice you think is correct. These are multiple-choice problems where you need to choose one answer from the list of possible answers.
>
> Look at the first example.
> It asks: 'What grade are you in? The correct answer choice is 1. Notice how the bubble beneath (below) the 1 has been shaded in for you. For some problems, you are going to mark your answer choices the same way, by shading in the bubble beneath (below) the answer choice you think is correct.
>
> For some problems, you will write the answer that you think is correct in a box.

Look at the second example. It says: 'Write the number four in the box.' The correct answer is written for you in the box. For some problems, you are going to write your answer the same way, by writing the answer you think is correct in a box."

Read the answers carefully. If you are not sure which answer is correct, mark the answer that you think is best. Make sure you mark an answer for all questions.

I will read all of the problems to you. Please do not say any answers out loud. You will answer all of the questions by writing on your paper.

You may underline words in the problems if you find that helpful. Also, feel free to use the white space on the assessment to work out your answers.

Are there any questions?

Address any questions.

If there are no more questions, turn to the page with the ▢ .

Pause; check to ensure all students are on the correct page.

I am going to read the problem one more time:

When you finish, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the dog at the top.

Pause; check to ensure all students are on the correct page.

I am going to read the problem one more time:

When you finish, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the frog at the top.

Pause; check to ensure all students are on the correct page.

I am going to read the problem one more time.

When you finish, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the balloons at the top.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the book at the top.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the car at the top.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the movie ticket at the top.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the soccer ball.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the smiley face.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the zebra.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the fish.

Pause; check to ensure all students are on the correct page.

Please complete the following problems on this page <u>and</u> the next page. Please write the correct answer in the box. When I say "begin'" you can start answering the questions. I will say end when time is up. Any questions?

Address any questions.

BEGIN.

Circulate as students work on the problems.
Provide students with ample time to complete the problems.

END.
Place your pencils down.

This concludes testing. Please sit quietly while I retrieve all testing materials.

Collect all testing materials.

Pre-test Student Information Sheet

INSTRUCTION: Please enter the information at the top of this form and provide the following information for ONLY those students in your class who have returned a signed Primary Grade Math Study Parent Consent Form that indicated parental consent to participate in the study. For each student, provide his or her unique district ID #, first and last name, indication of whether a completed Pre-test is enclosed, and any other relevant notes. Notes are optional; all other information is required.

School Name:		Testing Date:	
Teacher Name:		Testing Start Time:	
Grade Level(s):		Testing End Time:	

Were mathematics manipulatives used by students during the pre-test? (circle one)   YES   or   NO

Student's District ID #	Student's First Name	Student's Last Name	Completed Pre-test Enclosed (circle one)	Notes
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	

Student's District ID #	Student's First Name	Student's Last Name	Completed Pre-test Enclosed (circle one)	Notes
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	

Appendix D—Second Grade Administration Guide

Primary Grades Math Study:

Pre-test Guidelines, Administration Instructions, and Student Information Sheet

Grade 2

2013- 2014

Table of Contents

Pre-test Guidelines

Overview

The Primary Grades Math Study pre-test (hereafter, pre-test) provides four sections of assessments: Counting, Writing Numerals, Solving Word Problems and Addition/Subtraction Number Facts (First Grade).

The following guidelines provide information on the protocol for administering the pre-test. Throughout this document a second-person voice is used with the intended reader being the classroom teacher. It is assumed that the classroom teacher will administer the pre-test; however, it is permissible for other school personnel (such as a paraprofessional or even a substitute teacher) to administer the pre-test, providing they follow the pre-test protocol as detailed below.

Pre-test Testing Window

Please identify your locale in the table below for the applicable testing window.

Local Education Agency	Testing Window
District A	August 19 – August 30, 2013
District B	August 12 – August 23, 2013

Materials

The following materials are <u>required</u> for testing:

- Primary Grades Math Study Pre-test Guidelines and Administration Instructions (provided)
- A test booklet for each student (provided)
- At least one sharpened pencil for each student

The following materials are <u>encouraged</u> for testing:

- Counters and/or linking cubes for each student

Test Booklets

Test booklets are consumable and students mark their answers directly in the test booklets. Should you need additional testing materials, please contact Kristopher Childs (kristopher.childs@ucf.edu). Remember that these materials are to remain at the school site until the testing window has ended. The materials must be stored in a secure, access-restricted location at all times.

Students to be Tested

The pre-test for the Primary Grades Math Study will be administered to students who have returned signed consent forms indicating parental consent to participate in the study. On the pre-test student information sheet (p. 13 of this document), please list only those students for whom you have signed consent and provide their information in the table as requested. Only pre-tests completed by these students are to be relayed to project personnel.

At your discretion, the pre-test may also be administered to students who have not returned consent forms, with the understanding that students may return consent forms after the pre-test has been administered. In such a case, please retain possession of those students' pre-tests until such time that it is certain that parental consent is not granted.

Also at your school's discretion, the pre-test may be administered to students whose parents have declined consent to participate, so long as you do not relay their materials or data to project personnel. That is, you are free to use this test like you would any other test to assess your students' mathematics ability. Accordingly, it may make most sense to administer it to your whole class, irrespective of their status with the study—understanding that you will only relay materials for students who have signed parental consent.

Preparing for Testing

The first page of each test booklet has the following box for student information:

District:
School:
Teacher: Grade:
Student:

Prior to the testing session, the classroom teacher must enter this information (district name, school name, teacher name, student name, and student grade level) on each test booklet for each student to be tested. (Please do not leave it for students to enter this information.)

The pre-test for Primary Grades Math Study may be administered to students on either an individual or group basis. Please adhere to the following guidelines:

1. Ensure all students have testing materials (i.e., test booklet and a sharpened pencil).
2. Ensure that students and pre-labeled test booklets are properly paired (i.e., each student receives the test booklet that has his or her name written on it).
3. Provide students with a comfortable testing environment.
4. Testing administrators should adhere to the pre-test guidelines and administration Instructions.
5. No talking or communication between students is permitted during testing.
6. Students are permitted to use mathematics manipulatives during the pre-test.

Manipulatives

If students would ordinarily be permitted to use manipulatives in your classroom to solve math problems, then they should also be permitted for the pre-test.

Administering the Test

The testing conditions for the pre-test should be consistent with the testing conditions for other student assessments administered in the classroom. For example, students should space out the desks or use student "privacy folders" if that is what they would usually do.

Avoid reading problems or answering student questions in a way that may offer clues to the correct answer. Student responses should reflect their current math knowledge. Thus, it is important that effort is taken to ensure that the test questions are clearly presented and that students understand how they are to mark their answer; however, great care should be taken to not lead students to the correct answer. To ensure that the students' test responses are valid, it is important that appropriate procedures are followed when administering the pre-test. These procedures include:

- Administration of the appropriate test level (Grade 1 pre-test for Grade 1 students, etc.)

- Adherence to the pre-test guidelines and administration instructions in order to provide a standardized testing protocol across classrooms
- Maintenance of test security

Accommodations

Students with special academic plans (e.g., IEP, 504, ELL) may receive whatever accommodations are specified in their plans, at the teacher's discretion.

Testing in the Primary Grades

It is understood that children at this age level vary in their familiarity with whole-group testing procedures. The following recommendations are provided to facilitate a smooth testing procedure and minimize student frustration:

- Ensure students understand the testing instructions.
- Monitor students to ensure they are completing the correct question.
- Provide students with sufficient time to answer the questions.

When Students Get Stuck on a Problem

The following are suggested solutions for when students appear stuck and do not mark [or write] an answer for a given problem. Start with the first suggestion and only go on to subsequent suggestions if the prior ones did not resolve the student's delay in marking an answer:

1) Ask the student(s), "Would you like me to read the problem again?" Re-read the problem and accompanying directions if requested to do so.
2) Ask the student(s), "Do you have a question about how to mark your answer?" If the student answers in the affirmative, reiterate the directions from the first page on how to fill in the bubble or write in the box; whichever is appropriate for the given problem.
3) State, "I'm going to wait for another minute before going on to the next problem. Please look at the problem and mark [or write] what you think is a correct answer to that problem."
4) After waiting another minute, restate the direction to mark the answer for that given problem (for example, "Fill in the bubble that goes with your answer"), then read from the top of the script box for the next problem.
5) Tell the student it is okay if he or she skips that problem for now. He or she can come back to it after finishing the rest of the problems – if there is time.

Testing Time Allocation

Administration of the pre-test should take approximately 45 minutes. This is not a timed test, and students should be allowed adequate time to answer the test questions.

Submitting the Pre-test Materials

Upon conclusion of testing, separate out the test booklets for those students who have returned signed parental consent for participation in the study and repack them in the original packaging. Please be sure to include the pre-test guidelines, administration instructions, and completed student information sheet in the package. All unused test booklets should be repacked for return to project personnel. A Primary Grades Math Study representative will coordinate with your school to set a date to retrieve the testing materials from you. The target period of pickup will be the week of September 9.

Remaining test booklets will be for either those students of whom their parents have declined consent or have yet to return consent forms at all. Please retain the test booklets from this latter group of students (i.e., have not returned the consent form), in the event that they do bring back signed parental consent over the coming days or weeks. At that time, you will transfer their test booklet to a Primary Grades Math Study representative. If you have questions about this process, contact kristopher.childs@ucf.edu. To maintain the security of the test, please dispose of the test booklets for students whose parents have declined consent.

Pre-test Administration Instructions – Grade 2

[The boxes contain the script that you will read to the student.]

Your class is about to take a short math assessment. You will need a pencil.

Verify that all students have a pencil.

I will now pass out the assessments. The assessments are already labeled with your names. When you receive the assessment, keep it face up, and do not turn any pages; we will all begin at the same time after I go over the instructions.

Ensure that students and pre-labeled test booklets are properly paired (i.e., each student receives the test booklet that has his or her name written on it).

State the following box only if manipulatives are being used during pre-test administration.

For the math assessment it may help you to use manipulatives. I have placed manipulatives [indicate location of manipulatives]. The manipulatives can be used at any time during testing.

The first page of the assessment gives the instructions and provides samples of how you will mark your answers.

For some problems you will fill in the bubble beneath (below) the answer choice you think is correct. These are multiple-choice problems where you need to choose one answer from the list of possible answers.

Look at the first example.
It asks: 'What grade are you in? The correct answer choice is 2. Notice how the bubble beneath (below) the 2 has been shaded in for you. For some problems, you are going to mark your answer choices the same way, by shading in the bubble beneath (below) the answer choice you think is correct.

For some problems, you will write the answer that you think is correct in a box.

Look at the second example. It says: 'Write the number four in the box.' The correct answer is written for you in the box. For some problems, you are going to write your answer the same way, by writing the answer you think is correct in a box."

Read the answers carefully. If you are not sure which answer is correct, mark the answer that you think is best. Make sure you mark an answer for all questions.

I will read all of the problems to you. Please do not say any answers out loud. You will answer all of the questions by writing on your paper.

You may underline words in the problems if you find that helpful. Also, feel free to use the white space on the assessment to work out your answers.

Are there any questions?

Address any questions.

Turn to the page with the dog at the top.

Pause; check to ensure all students are on the correct page.

I am going to read the problem one more time:

When you finish, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the frog at the top.

Pause; check to ensure all students are on the correct page.

I am going to read the problem one more time:

When you finish, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the car at the top.

Pause; check to ensure all students are on the correct page.

Write it in the box.
I am going to read the problem one more time.
Write it in the box.
When you finish, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the balloons at the top.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the book at the top.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the pencil at the top.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the movie ticket at the top.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the soccer ball.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the smiley face.

Pause; check to ensure all students are on the correct page.

Shade in the circle below the answer you think is correct.

I am going to read the problem one more time:

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the zebra.

Pause; check to ensure all students are on the correct page.

I am going to read the problem one more time:

Shade in the circle below the answer you think is correct.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the fish.

Pause; check to ensure all students are on the correct page.

Please complete the following problems on this page and the next page. Please write the correct answer in the box. When I say "begin," you can start answering the questions. I will say "end" when time is up. Any questions?

Address any questions.

BEGIN.

Circulate as students work on the problems.
Provide students with ample time to complete the problems.

END.
Place your pencils down.

This concludes testing. Please sit quietly while I retrieve all testing materials.

Collect all testing materials.

Pre-test Student Information Sheet

INSTRUCTION: Please enter the information at the top of this form and provide the following information for ONLY those students in your class who have returned a signed Primary Grade Math Study Parent Consent Form that indicated parental consent to participate in the study. For each student, provide his or her unique district ID #, first and last name, indication of whether a completed Pre-test is enclosed, and any other relevant notes. Notes are optional; all other information is required.

School Name:		Testing Date:	
Teacher Name:		**Testing Start Time:**	
Grade Level(s):		**Testing End Time:**	

Were mathematics manipulatives used by students during the pre-test? (circle one)   YES   or   NO

Student's District ID #	Student's First Name	Student's Last Name	Completed Pre-test Enclosed (circle one)	Notes
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	

Student's District ID #	Student's First Name	Student's Last Name	Completed Pre-test Enclosed (circle one)	Notes
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	
			YES or NO	

Appendix E—Distributions of Number of Items Answered Correctly Within Each Factor

Figure 12. Distribution of the number of items individual students in the grade 1 sample answered correctly within the Counting factor.

Figure 13. Distribution of the number of items individual students in the grade 1 sample answered correctly within the Word Problems factor.

Figure 14. Distribution of the number of items individual students in the grade 1 sample answered correctly within the Computation factor.

Figure 15. Distribution of the number of items individual students in the grade 2 sample answered correctly within the Counting factor.

Figure 16. Distribution of the number of items individual students in the grade 2 sample answered correctly within the Word Problems factor.
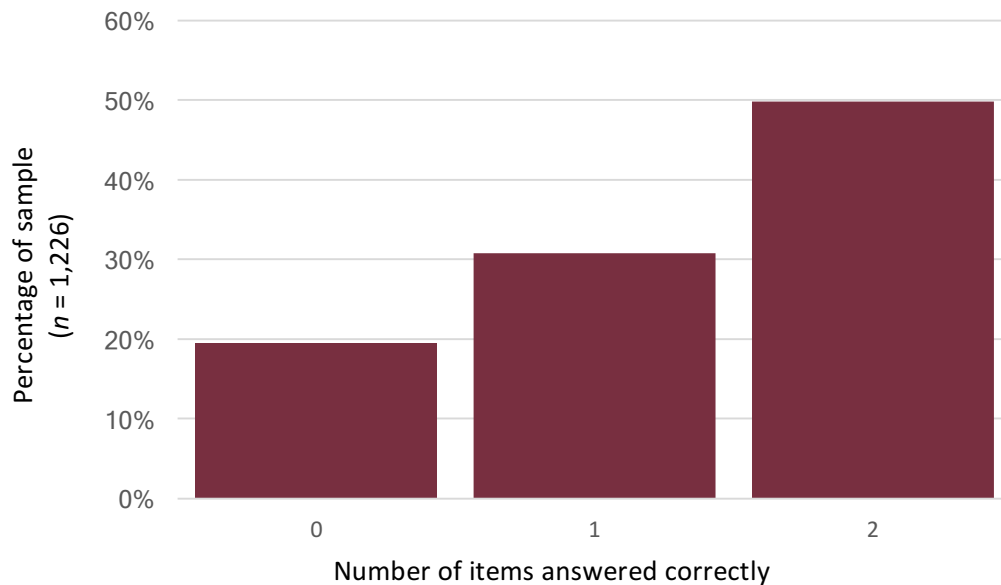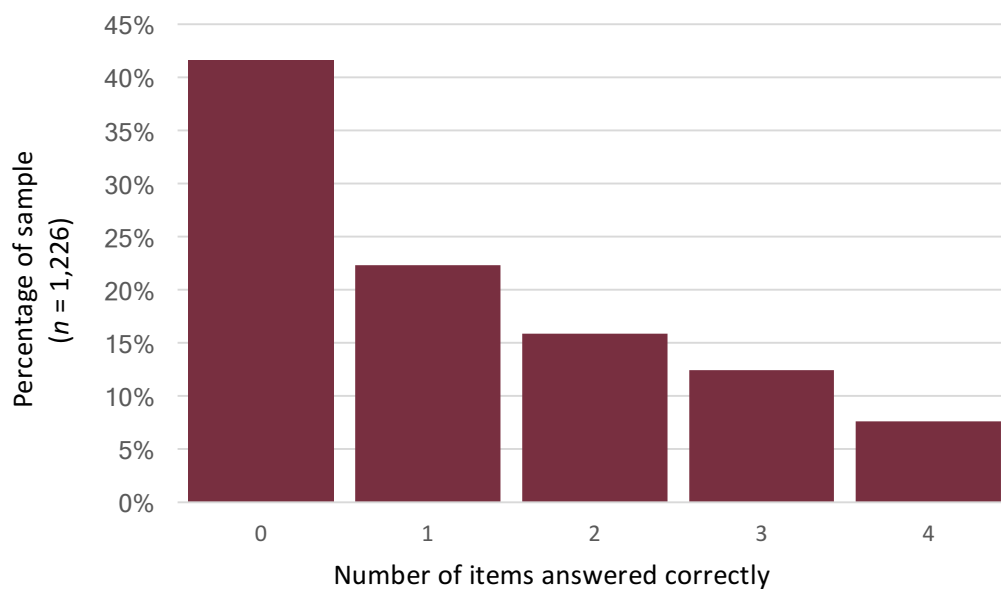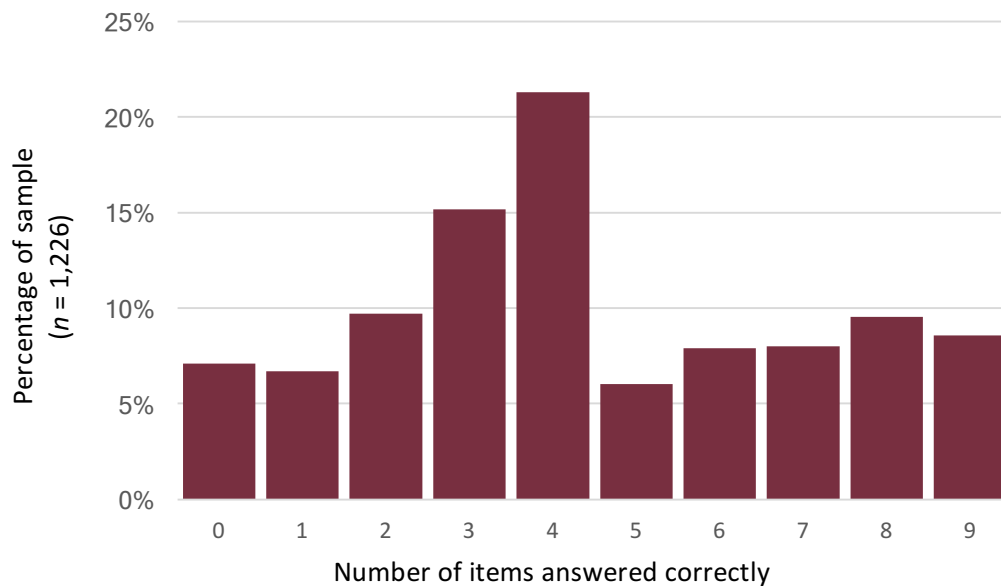

Figure 17. Distribution of the number of items individual students in the grade 2 sample answered correctly within the Computation factor.
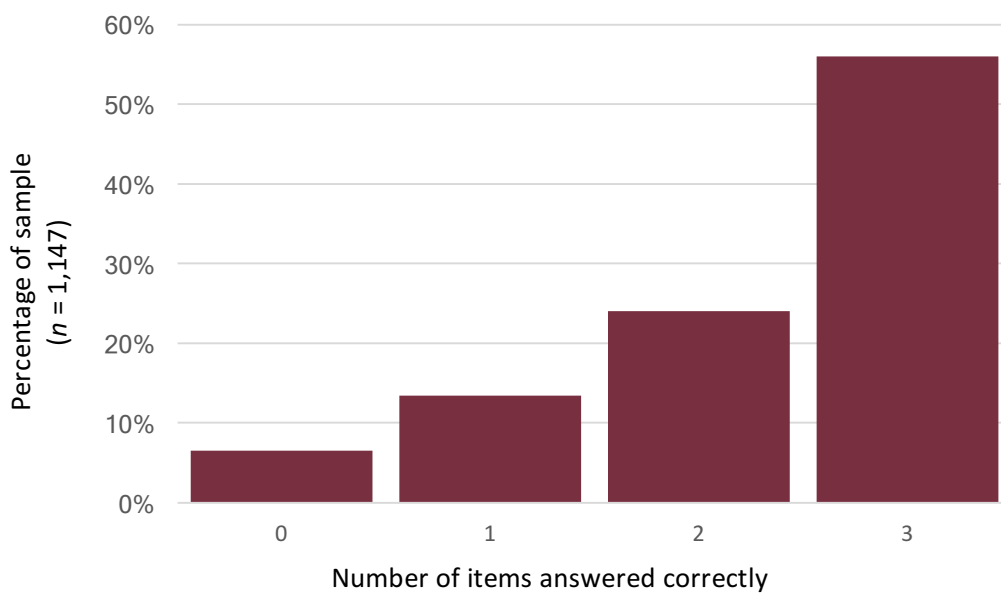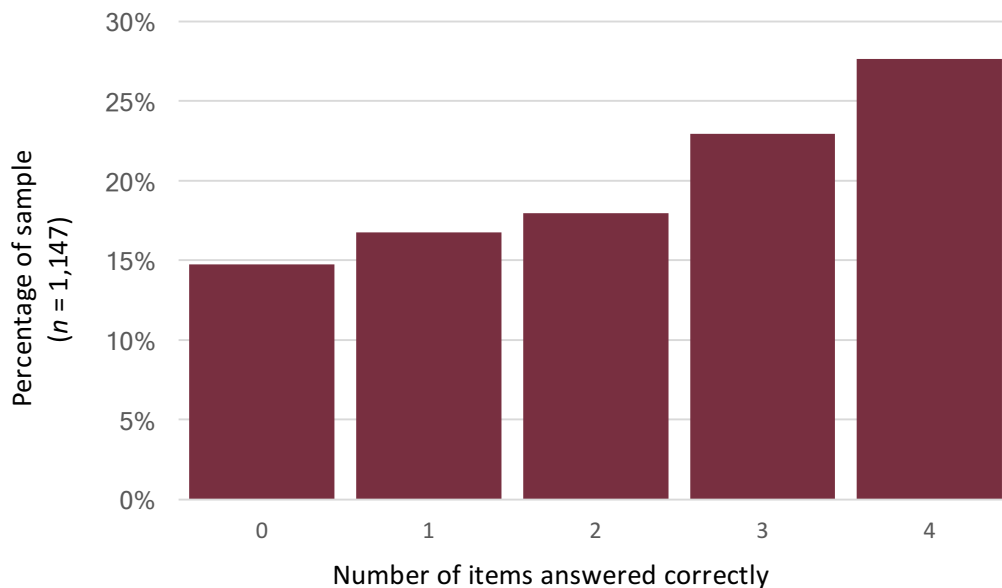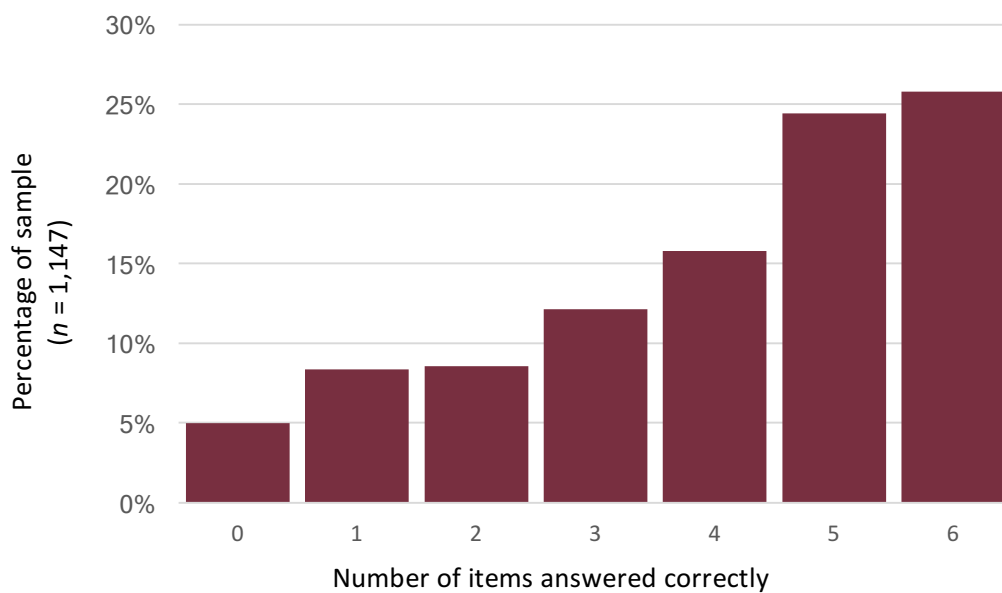
Appendix F—Most Common Incorrect Response for Each Item

Table 19. Proportion of Grade 1 Responses by Item

Item	Item description	Correct response Response (%)	Most frequent incorrect responses Response (%)	Response (%)	Response (%)	Response (%)
Counting						
1		7 (.97)	8 (.01)	6 (.01)	NA (<.01)	UI (<.01)
2		7 (.76)	8 (.05)	10 (.03)	1 (.03)	6 (.02)
3		13 (.54)	15 (.10)	1 (.06)	31 (.02)	14 (.02)
Word Problems						
4		7 (.79)	6 (.07)	1 (.06)	3 (.06)	4 (.02)
5		2 (.34)	6 (.31)	10 (.24)	8 (.05)	4 (.05)
6		12 (.40)	7 (.48)	4 (.08)	3 (.02)	NA (.01)
7		4 (.17)	7 (.41)	10 (.30)	3 (.05)	21 (.05)
8		10 (.58)	7 (.16)	24 (.08)	1 (.08)	17 (.07)
9		5 (.35)	14 (.23)	23 (.18)	9 (.14)	6 (.08)
10		5 (.36)	11 (.24)	6 (.15)	17 (.12)	16 (.11)
Computation						
11		11 (.67)	10 (.12)	6 (.04)	7 (.04)	12 (.02)
12		3 (.39)	9 (.34)	8 (.06)	4 (.04)	6 (.03)
13		7 (.79)	8 (.04)	6 (.04)	NA (.03)	5 (.02)
14		3 (.32)	17 (.28)	7 (.04)	10 (.03)	8 (.03)
15		6 (.78)	5 (.03)	NA (.03)	7 (.03)	2 (.02)
16		4 (.34)	10 (.37)	5 (.04)	9 (.04)	7 (.04)
17		6 (.28)	18 (.18)	NA (.06)	4 (.05)	16 (.05)
18		11 (.25)	19 (.21)	NA (.08)	10 (.05)	12 (.03)
19		16 (.40)	17 (.10)	NA (.07)	15 (.05)	8 (.04)
20		8 (.61)	7 (.08)	NA (.07)	6 (.06)	3 (.03)

Note. *n* = 1,226 valid grade 1 tests conducted. Items that remain in models after factor analysis are presented in boldface type. Only the four most common incorrect responses are displayed. Percentages may not sum to 100. Items that were not answered were recorded as "NA". Item responses that were unclear were recorded as "UI".

Table 20. Proportion of Grade 2 Responses by Item

Item	Item description	Correct response Response (%)	Most frequent incorrect responses Response (%)	Response (%)	Response (%)	Response (%)
Counting						
1		15 (.89)	16 (.02)	14 (.01)	13 (.01)	20 (.01)
2		49 (.73)	14 (.07)	40 (.04)	51 (.02)	60 (.01)
3		92 (.68)	90 (.05)	91 (.04)	83 (.04)	93 (.02)
Word Problems						
4		15 (.89)	17 (.05)	1 (.04)	8 (.01)	7 (.01)
5		6 (.53)	40 (.20)	23 (.15)	7 (.06)	17 (.05)
6		24 (.56)	10 (.27)	16 (.12)	4 (.02)	6 (.02)
7		3 (.74)	11 (.10)	7 (.08)	4 (.05)	28 (.20)
8		9 (.74)	8 (.12)	11 (.10)	17 (.02)	25 (.02)
9		4 (.49)	3 (.26)	9 (.12)	15 (.08)	12 (.04)
10		7 (.67)	6 (.14)	13 (.11)	33 (.04)	20 (.11)
Computation						
11		11 (.92)	10 (.20)	12 (.02)	9 (.01)	6 (<.01)
12		18 (.82)	10 (.05)	19 (.03)	17 (.02)	16 (.01)
13		18 (.84)	17 (.03)	19 (.02)	16 (.01)	8 (.01)
14		3 (.76)	17 (.13)	4 (.02)	2 (.02)	7 (.01)
15		26 (.68)	27 (.05)	25 (.05)	16 (.02)	8 (.02)
16		6 (.66)	18 (.14)	5 (.04)	7 (.03)	4 (.03)
17		6 (.64)	24 (.08)	5 (.07)	7 (.05)	NA (.02)
18		30 (.59)	90 (.04)	NA (.04)	20 (.04)	31 (.03)
19		42 (.43)	10 (.07)	41 (.05)	36 (.05)	NA (.05)
20		6 (.53)	5 (.08)	28 (.08)	NA (.05)	7 (.04)

Note. n = 1,147 valid grade 1 tests conducted. Items that remain in models after factor analysis are presented in boldface type. Only the four most common incorrect responses are displayed. Percentages may not sum to 100. Items that were not answered were recorded as "NA". Item responses that were unclear were recorded as "UI".