Florida State University Libraries

2017

Elementary Mathematics Student Assessment: Measuring the Performance of Grade K, 1, and 2 Students in Number, Operations, and Equality in Spring 2016

Robert C. Schoen, Daniel Anderson and Charity Bauduin

Elementary Mathematics Student Assessment

Measuring the Performance of Grade K, 1, and 2 Students in Number, Operations, and Equality in Spring 2016

Robert C. Schoen
Daniel Anderson
Charity Bauduin

Suggested citation: Schoen, R. C., Anderson, D., & Bauduin, C. (2017). *Elementary mathematics student assessment: Measuring the performance of grade K, 1, and 2 students in number, operations, and equality in spring 2016*. (Research Report No. 2017-22). Tallahassee, FL: Learning Systems Institute, Florida State University. http://doi.org/10.17125/fsu.1534964774

Detailed information about items are not included in this report. This information was removed in order to release the psychometric report and maintain test security. Requests to view the full report should be directed to Robert Schoen (rschoen@lsi.fsu.edu).

Elementary Mathematics Student Assessment

Measuring the Performance of Grade K, 1, and 2 Students in Number, Operations, and Equality in Spring 2016

Robert C. Schoen[1]

Daniel Anderson[2]

Charity Bauduin[1]

May 2017

[1]Florida State University

[2]University of Oregon

Acknowledgments

The successful development and implementation of this assessment involved many experts in mathematics education and many more students. Some of the key people involved in development of the test are listed here along with their roles in the endeavor.

Robert Schoen designed the content and format of the test, coordinated the feasibility study, created the scoring criteria, interpreted the results, and oversaw the writing of this report. Daniel Anderson performed the data analysis for the item calibration, exploratory factor analysis, item-response theory-based models, and vertical linking between grade levels, and he also contributed to writing the report. Charity Bauduin reviewed the alignment of items with the Mathematics Florida Standards, managed the report-writing process, and assisted with editing the style and format of the report.

Zachary Champagne assisted with the development and production of the test. Amanda Tazaz coordinated the dissemination and collection of the tests and corresponding consent forms. Kristy Farina designed and managed the data-entry system, trained data-entry personnel on the system, verified the accuracy of the data, and assisted with description of the data-entry process and sample descriptives for the present report. Alexandra Utecht, Shelby McCrackin, Senai Tazaz, and Claire Riddell served as data-entry personnel. Claire Riddell, Mark McClure, and Monica Hurdal served as reviewers of the test items, response options, and scoring. Anne B. Thistle provided valuable assistance with copy editing.

We are especially grateful for the support from the Math-Science Partnership grant program and the Florida Department of Education and to the students, parents, principals, district leaders, and teachers who agreed to participate in the study.

Table of Contents

List of Appendices

List of Tables

List of Figures

List of Equations

List of Abbreviations

2PL .... Two-parameter logistic

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

CDU .... Compare Difference Unknown

CTT .... Classical Test Theory

ELL .... English Language Learner

EMSA .... Elementary Mathematics Student Assessment

IRT .... Item Response Theory

JCU .... Join Change Unknown

JRU .... Join Result Unknown

JSU .... Join Start Unknown

MAP .... Minimal Average Partial

MD .... Measurement Division

MG .... Multiplication Grouping

MPAC .... Mathematics Performance and Cognition

NEAT .... Non-Equivalent Groups With Anchor Test

PWU .... Part-Part-Whole Whole Unknown

SWD .... Students With Disabilities

VSS .... Very Simple Structure

Executive Summary

The present report describes the Elementary Mathematics Student Assessment (EMSA) as it was used with grade K, 1, and 2 students in spring 2016. Although several versions of EMSA are available, each designed for different purposes, we will refer to the set of forms discussed here as the Spring 2016 K–2 EMSA.

The Spring 2016 K–2 EMSA was designed to serve as a mathematics achievement test administered to students at the end of the school year. It contained items involving number, word problems, basic addition and subtraction facts, addition and subtraction involving multidigit numbers, and the meaning of the equals sign in mathematics.

Purpose

The intended use of the test was to serve as an outcome measure of student achievement in a randomized controlled trial evaluating the impact of a teacher professional-development program called Cognitively Guided Instruction on mathematics students. The purpose of the current report is to serve as a reference document that describes the content of the test, the development process, the administration protocol, the scoring procedures, and the process we used to create the final scale for the Spring 2016 K–2 EMSA.

Our primary motivation is to create a reference document that details the development/validation process that we undertook and to archive the results of that work to serve as a reference. The goal of the report is therefore to support transparency in our research so that scrutiny can be duly applied by the research community and to allow peers and colleagues to provide critical feedback. The intended audience is educational researchers and program evaluators who may be interested in using the instrument in the future. Should the opportunity arise, we intend to use this report to provide sufficient information for ourselves or others to be able to replicate the administration and scoring of the data in the future. We hope a secondary benefit will be to provide those conducting similar investigations an opportunity to benefit from the findings and lessons we learned through the worked reported here.

Content

In general, the test was designed to align with the core content in the number and operations domains in the Common Core State Standards for Mathematics (CCSS-M; NGACBP & CCSSO, 2010) in grades K, 1, and 2. In a few cases, the content of the items goes beyond strictly interpreted content limits implied by the CCSS-M. For example, some equals-sign items involve more than three quantities (e.g., $a + b = c + d$) in grade 1, and items on the grades 1 and 2 involve grouping-type problems as a way to test students' problem-solving abilities as well as their understanding of the base-ten number system. The current report focuses on the overall scale, but the equals-sign items were used to create an additional scale that was focused specifically on student understanding of the equals sign as a relational operator. Details about the equals-sign scale will be provided in an addendum to the current report.

Sample and Setting

The 2016 Spring EMSA tests were administered to 4,535 participating grade K, 1, and 2 students in 66 schools located in 9 public school districts in Florida during spring 2016. The sample included 950 grade K, 1,821 grade 1, and 1,764 grade 2 students. The school districts were primarily using *GoMath!* (Dixon, Larson, Leiva, & Adams, 2013), a curriculum series designed to be aligned with the Mathematics Florida

Standards (Florida Department of Education, 2014), which are very similar to the CCSS-M (NGACBP & CCSSO, 2010).

Test Specifications and Administration

The Spring 2016 K–2 EMSA includes selected-response and constructed-response test items at each grade level. On selected-response items, students were asked to mark their answer choices by filling in the bubble beneath the answer choice they think is correct. Selected-response options are based on common responses students at these grade levels have provided in previous versions of these items presented in a constructed-response format in cognitive interviews and field-tests of paper-pencil tests. The response options were presented horizontally; the five response options are sequenced left to right with numbers from least to greatest. On constructed-response items, students are asked to write their answer choices in the box provided. The grade K test form consists of 21 items, the grade 1 test 26 items, and the grade 2 test 25 items. After reviewing item statistics on the basis of classical test theory, item-response theory model calibration, and vertical scaling, the final scale for the grades K, 1, and 2 tests draw on 18, 23, and 23 items, respectively.

Data Entry and Scoring

Research assistants performed data entry by typing data from each page of every student test into a database using FileMaker Pro software (Version 14.1). Student responses were recorded by means of dropdown menus. A double-entry procedure was performed with a random sample of 11% of the tests. The first and second entries matched on more than 99% of the item-level responses that were entered.

Although each item on the Spring 2016 EMSA was designed to have a unique, correct solution, students could generate equivalent responses (e.g., 5, 4 + 1). To accommodate this possibility, an adjudication committee met to review the set of all responses to each item and determine the set of correct answers.

Overall test scores were determined by means of a two-parameter logistic model based on item-response theory; the discrimination parameter was freely estimated for the majority of items but constrained to be equivalent for all items specifically designed to measure student understanding of the mathematical meaning of the equals sign. The student ability estimates were mapped onto a single scale by the Stocking-Lord method for vertical equating (Kolen & Brennan, 2014).

Reliability

At grade K, the distribution of item difficulties matched the distribution of person abilities quite well; as a result, the vast majority of the person-ability distribution was within the reliable range. At grades 1 and 2, the person-ability distribution was shifted slightly to the right of the item-difficulty distribution, implying that a significant proportion of students had a higher ability than the estimated difficulty of the items. This difference was also evident in the distribution of total raw scores on the grade 1 and 2 tests, where 5 and 7% of the sample students, respectively, responded to every item on the test correctly. The upper tail of the person distribution was therefore outside of the desired reliability range for the grade 1 and 2 samples.

Summary

The content of the Spring 2016 K–2 EMSA test aligns with the grade-level expectations in the CCSS-M and the Mathematics Florida Standards in the domains of Number and Base Ten as well as Operations and Algebraic Thinking. The tests were recognized by teachers as being relevant to what they teach, and they were able to administer the tests within the usual constraints of the school day.

The test difficulty aligned with student abilities at grade K very well, but results of data analysis suggest the test may not discriminate as well among higher-achieving students in grades 1 and 2. Better alignment between the overall difficulty of the test and student abilities in grades 1 and 2 may result in increased reliability, especially for student ability levels in the upper tails of the distribution at those grade levels. The test is not designed to discriminate among individual students, but this apparent limitation may dampen the ability of the Spring 2016 K–2 EMSA to serve its primary purpose, which was to detect potential achievement differences between students in treatment and control schools in a randomized trial. The comparison of student achievement may be most valid in the grade K subsample and less so in the grade 1 and 2 subsamples.

1. Introduction and Overview

The Spring 2016 K–2 EMSA test forms represent an extension of our work in the development and implementation of the Fall 2013, Fall 2014, and Fall 2015 EMSA tests (Schoen, Anderson, Champagne, & Bauduin, 2017; Schoen, LaVenia, Bauduin, & Farina, 2016a, 2016b) and the spring 2014 and spring 2015 Mathematics Performance and Cognition (MPAC) Interviews (Schoen, LaVenia, Champagne, & Farina, 2016c; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016d). The Spring 2016 K–2 EMSA contains three test forms, corresponding to grades K, 1, and 2. Like the Fall 2015 K–2 EMSA, the Spring 2016 K–2 EMSA is scored on a single, vertically equated scale, but the tests are not equated across seasons (i.e., fall, spring) or years (e.g., 2015, 2016).

The Spring 2016 K–2 EMSA was designed to serve as a mathematics achievement test administered to students at the end of the school year and was intended to measure students' ability to solve problems involving number, operations, and equality. It did not attempt to measure mathematics knowledge or ability in other domains such as geometry, measurement, probability, or data analysis. The test scores were used as an outcome measure for a randomized trial investigating the effect of a teacher professional-development program on student mathematics ability.

The Spring 2016 K–2 EMSA test forms were used to create a vertically scaled score, by means of item-response theory, that is directly comparable across the three grades. The vertically scaled score increases statistical power in the randomized trial by allowing the data to be pooled across grade levels, effectively tripling the sample size over those of treatment-control comparisons within each grade level.

The K–2 EMSA tests were designed to be administered in a whole-group setting in a paper-pencil format. Test administrators are given a guide explaining how to administer the tests, along with a script to use while doing so. Questions were read aloud to students, and students shaded bubbles to indicate their responses to multiple-choice items or wrote their responses to constructed-response items in boxes. Test administrators are encouraged to allow students to use manipulatives in accordance with their typical classroom practice.

The current report focuses on the content of the test, the development process, the administration protocol, the scoring procedures, and the process we used to create the final scale for the Spring 2016 K–2 EMSA. Its purpose is to serve as a reference document that describes available evidence to support the substantive, structural, and external validity arguments (Flake, Pek, & Hehman, 2017) and the process we used to create the final scale. Although these elements may provide valuable information to other researchers, they also serve as a reference upon which we can base continual future improvement of our design and field-testing of assessment instruments.

The current report focuses on the overall test and resulting ability estimates on the overall scale. The items designed specifically to measure student understanding of the equals sign as a relational operator were also used to create an additional scale. More information on that scale will be provided in a future report, which will be written as an addendum to this one.

The second chapter of the report describes the test-development process and the alignment of the content of the test with current mainstream curriculum standards in place for grade K, grade 1, and grade 2 students in mathematics. It describes the test and item specifications as well as the administration instructions, scoring protocol, and data-management procedures. The actual test booklets used by students are provided in Appendices A, B, and C, and the administration instructions are provided in Appendices D, E, and F.

The third chapter describes the data-analytic procedures used, ultimately, to generate the final scale and scores from the Spring 2016 EMSA. The first steps in the analytic process involved initial screening of the test items by means of statistical techniques based on classical test theory (CTT; Crocker & Algina, 2008). Items with particularly poor statistics were reviewed by content experts, who determined whether to remove them from the scale. Next steps involved an analysis of the dimensionality of the test by means of exploratory factor analysis and data modeling based on item response theory (IRT) that used two-parameter logistic (2PL) models, separately for each grade level.

The results of the screening and scaling process as well as information about scale reliability are presented in chapter four. The fifth chapter provides a discussion and reflection on the findings as well as recommendations for improvement of the test and other potential next steps.

1.1. Test Overview

Table 1.1 provides an overall blueprint for each of the three tests.

Table 1.1. Final Blueprint for the Spring 2016 K–2 EMSA Test

	Number of items		
Section	Grade K	Grade 1	Grade 2
Number Facts	5	5	4
Operations on Both Sides of the Equals Sign	4	5	5
Word Problems	4	5	5
Equals Sign as a Relational Symbol	5	5	5
Computation	3	6	6
Total	21	26	25

By design, within each section on test forms at adjacent grade levels, at least three items were identical, to permit vertical scaling across grade levels. For the most part, when the questions were not identical, those for the upper grades were similar in nature but involved higher numbers and were therefore proportionally more difficult. The higher numbers were also intended to reveal information about how these older students made sense of operations on multidigit whole numbers. In general, the items were intended to be sequenced from easier to more difficult within each subsection.

During test administration, students recorded their responses with a pencil directly on the test booklet. Students were allowed to use blank space provided in the test booklet to determine their answers. In most cases, the students' classroom teacher administered the test.

The test administrators were instructed to read each problem in the Number Facts, Word Problems, and Equals Sign as a Relational Symbol sections aloud twice to students. The administrator was given the flexibility to reread a problem on request but was instructed always to read the entire problem exactly as written and to refrain from reading just a portion of it. After the problems in these sections were read by the administrator, the students were instructed to "fill in the bubble under the correct answer" or to "circle yes if the equation is correct or no if the equation is not correct," depending on the item format. Almost all items in the Computation section and several items in the Operations on Both Sides of the Equals Sign used a constructed-response format. Students completed these items independently and were instructed to write the number in the box that would make the equation correct.

The testing conditions for the Spring 2016 K–2 EMSA were expected to be held consistent with the testing conditions used in other student assessments administered in the teacher's classroom. For example, students should separate their desks or use student "privacy folders" if that is what they usually do. In addition, students were permitted to use mathematics manipulatives during the Spring 2016 K–2 EMSA if they were ordinarily permitted to do so in that particular classroom.

The Spring 2016 K–2 EMSA test was not timed, and sufficient time was provided for students to solve each problem. Test administrators were informed that the test required approximately 45 minutes to administer, but administration time was allowed to vary across settings. Test administrators were encouraged to provide students with sufficient time to complete each item on the test to their own satisfaction.

1.1.1. Number Facts Section

Table 1.2 provides an overview of the Number Facts items by grade level. The anchor set for the grades K–1 includes three items, that for the grades 1–2 includes four. Two items in this section (▨ = □ and ▨ = □) were included at all three grade levels, so as to create a set of anchor items linking all three grade levels.

Table 1.2. Items in the Number Facts Section

Variable name	Response format	Grade K	Grade 1	Grade 2
GKi16	CR	▨ = □		
GKiG1i17	CR	▨ = □	▨ = □	
GKG1i18_G2i15	CR	▨ = □	▨ = □	▨ = □
GKi19	CR	▨ = □		
GKi20_G1i19_G2i16	CR	▨ = □	▨ = □	▨ = □
G1i20_G2i17	CR		▨ = □	▨ = □
G1i21_G2i18	CR		▨ ▨ = ▨	▨ ▨ = ▨

Note. CR = constructed response.

All of the items in this section at each grade level were grouped with the computation items and were therefore not read aloud to students. Students worked at their own pace to solve and wrote their answers in the box provided for each problem.

1.1.2. Operations on Both Sides of the Equals Sign Section

Table 1.3 provides an overview of the Operations on Both Sides of the Equals Sign items by grade level. The anchor set for the grades K–1 includes four items, that for the grades 1–2 includes five. Four items in this section (▨ , ▨ , ▨ , and ▨ ▨ ▨) were included at all three grade levels, so as to create a set of anchor items among the three grade levels.

Table 1.3. Items in the Operations on Both Sides of the Equals Sign Section
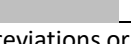
Variable name	Response format	Grade K	Grade 1	Grade 2
GKG1G2i8	SR			
GKG1G2i10	SR			
GKG1G2i13	SR			
GKi22_G1i25_G2i23	CR			
G1i26_G2i24	CR			

Note. CR = constructed response; SR = selected-response.

1.1.3. Word Problems Section

The Word Problems section contained a set of word problems representing a range of difficulty and two subtypes: (1) standard addition and subtraction and (2) standard multiplication and division (grouping- and measurement-type problems, respectively).

Table 1.4. Items in the Word Problems Section

Variable name	Response format	Grade K	Grade 1	Grade 2
GKi2	SR			
GKi4_G1i2	SR			
GKi3_G1i1	SR			
GKi5_G1i3_G2i1	SR			
G1i4_G2i2	SR			
G1i5_G2i3	SR			
G2i4	SR			
G2i5	SR			

Note. SR = selected-response. For a full list of the problem-type abbreviations, see the List of Abbreviations or Carpenter et al. (2015).

Table 1.4 shows the types of problems included in the Word Problems section of the Spring 2016 K–2 EMSA test at each grade level. The three-letter abbreviations for the problem types correspond to the names of word problems as defined by Carpenter, Fennema, Franke, Levi, & Empson (2015). The numbers correspond to the two numbers given in the problem.

As indicated in Table 1.4, three identical items make up the grades K–1 anchor set and three identical items the grades 1–2 anchor set. One anchor item is used at all three grade levels in this section of the test.

The two *multiplication-grouping* problems and the *measurement-division* problem might be considered to be beyond the scope of the content of the CCSS-M at grades K, 1 and 2. We include them in the tests, because abundant empirical evidence demonstrates that students at these grade levels can solve these types of problems (Carpenter, Ansell, Franke, Fennema, & Weisbeck, 1993; Turner & Celedón-Pattichis, 2011; Verschaffel, Greer, & DeCorte, 2007). Moreover, the focus on place value and the base-ten structure of the number system in the mathematics curriculum standards at the early elementary level involves grouping situations—with a particular focus on groups of ten. This is consistent with *Multiplication-Grouping* and *Measurement-Division* problems (Carpenter et al., 2015).

All the items in this section at each grade level are read aloud twice to the students as per the administration instructions (see Appendices D, E, and F) to help focus the assessment on students' mathematics achievement and lessen the potential effect of their respective reading or listening comprehension ability.

1.1.4. Equals Sign as a Relational Symbol Section

Table 1.5 provides an overview of the items in the Equals Sign as a Relational Symbol section of the Spring 2016 K–2 EMSA test at each grade level. All five items in this section served as anchor items across all three grade levels to increase the probability of success in vertical scaling.

Table 1.5. Items in the Equals Sign as a Relational Symbol Section

Variable name	Response format	Grade K	Grade 1	Grade 2
GKG1G2i6	SR			
GKG1G2i7	SR			
GKG1G2i12	SR			
GKG1G2i9	SR			
GKG1G2i11	SR			

Note. SR = selected-response.

This section of the test is designed to measure students' understanding of the equals sign as a relational symbol. The equations in this section are read aloud to students. Students are asked to circle "yes" if the equation is correct, and to circle "no" if the equation is not correct.

1.1.5. Computation Section

Table 1.6 provides an overview of the items in the Computation section of the Spring 2016 K–2 EMSA test at each grade level. In this section, the grades K–1 anchor set included three items; the grades 1–2 anchor set also included three items. Because of the very wide range of ability in computation between grade K and grade 2 students, no items were identical across all three grade levels in this section.

Table 1.6. Items in the Computation Section

Variable name	Response format	Grade K	Grade 1	Grade 2
GKG1i14	SR			
GKG1i15	SR			
G1i16_G2i14	CR			
GKi21_G1i22	CR	$= \square$	$= \square$	
G1i23_G2i19	CR		$= \square$	$= \square$
G1i24_G2i20	CR		$= \square$	$= \square$
G2i21	CR			$= \square$
G2i22	CR			$= \square$
G2i25	CR			

Note. CR = constructed response; SR = selected-response.

The final section of the test was designed to measure students' ability to compute sums and differences with basic facts and higher numbers. The problems in this section were more varied, as second-grade
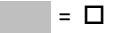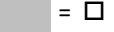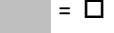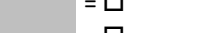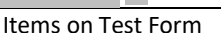
students are typically more proficient with computing sums and differences of greater numbers than kindergarten or first-grade students.

For this section, students were given directions for completing the items on their own. They were told that they would work on some addition and subtraction problems on their own, and they would solve them at their own pace. According to the script in the administration guide, the students were encouraged to look closely at the symbol to decide whether each problem involved addition or subtraction.

1.1.6. Detailed Test Blueprint

Table 1.7 provides a detailed blueprint showing the items in each of the five sections of the test (i.e., Number Facts, Operations on Both Sides of the Equals Sign, Word Problems, and Equals Sign as a Relational Symbol, and Computation). Items displayed with a strikethrough were on the test form but were removed from the final scale as a result of poor item statistics. See Chapter 3 of the present report for more information on the review and analysis of the individual items.

Table 1.7. Detailed Test Blueprint for the Spring 2016 K–2 EMSA

Item description	Variable names		
	Grade K	Grade 1	Grade 2
Number Facts			
= ☐	GKi16		
= ☐	GKiG1i17	GKiG1i17	
= ☐	GKG1i18_G2i15	GKG1i18_G2i15	GKG1i18_G2i15
= ☐	GKi19		
= ☐	GKi20_G1i19_G216	GKi20_G1i19_G216	GKi20_G1i19_G216
= ☐		G1i20_G2i17	G1i20_G2i17
		G1i21_G2i18	G1i21_G2i18
Operations on Both Sides of the Equals Sign			
	GKG1G2i8	GKG1G2i8	GKG1G2i8
	GKG1G2i10	GKG1G2i10	GKG1G2i10
	GKG1G2i13	GKG1G2i13	GKG1G2i13
	~~GKi22_G1i25_G2i23~~	GKi22_G1i25_G2i23	GKi22_G1i25_G2i23
		G1i26_G2i24	G1i26_G2i24
Word Problems			
	GKi2		
	GKi4_G1i2	GKi4_G1i2	
	GKi3_G1i1	GKi3_G1i1	
	GKi5_G1i3_G2i1	GKi5_G1i3_G2i1	GKi5_G1i3_G2i1
		G1i4_G2i2	G1i4_G2i2
		~~G1i5_G2i3~~	G1i5_G2i3
			G2i4
			G2i5
Equals Sign as a Relational Symbol			
	~~GKG1G2i6~~	~~GKG1G2i6~~	~~GKG1G2i6~~
	~~GKG1G2i7~~	~~GKG1G2i7~~	~~GKG1G2i7~~
	GKG1G2i12	GKG1G2i12	GKG1G2i12
	GKG1G2i9	GKG1G2i9	GKG1G2i9
	GKG1G2i11	GKG1G2i11	GKG1G2i11
Computation			
	GKG1i14	GKG1i14	
	GKG1i15	GKG1i15	
		G1i16_G2i14	G1i16_G2i14
= ☐	GKi21_G1i22	GKi21_G1i22	
= ☐		G1i23_G2i19	G1i23_G2i19
= ☐		G1i24_G2i20	G1i24_G2i20
= ☐			G2i21
			G2i22
			G2i25
Items on Test Form	21	26	25
Items in Final Scale	18	23	23

Note. The eight items with strikethrough font are the items that were on the test form but were removed from the final scale.

1.2. Test Administration

Teachers were given detailed instructions (which appear in Appendices D, E, and F) on how to administer the test, including a script to use during administration. Test administrators were asked to write students' names on the front covers of the tests to increase legibility and accuracy in data entry. They were also instructed to permit students to use manipulable materials if that was common practice in their classrooms. For the first part 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. For the last part of the test, teachers read aloud directions, and students solved equations individually at their own pace. Teachers were encouraged to provide appropriate testing accommodations for students, as necessary, in accordance with their individual educational plans.

Teachers were instructed to insert completed tests into an opaque sealed envelope and to deliver the envelope 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 and Setting

Students in the field-test sample attended schools where their teachers had volunteered to participate in a randomized controlled trial of a year-long professional-development program in mathematics called Cognitively Guided Instruction. Tests forms were delivered to schools by project staff during the week of April 11–15, 2016. In the field tests reported here, the students' classroom teachers administered the tests during the testing window of April 18–May 11.

The analytic sample included 4,535 students drawn from 271 classrooms in 9 Florida public school districts. Table 1.8 provides descriptive statistics for the data we have at the time of this report.

Table 1.8. Demographic Characteristics of the Students in the Spring 2016 Field Test of the K–2 EMSA Tests

Student characteristic	Number (proportion of sample or subsample)			
	Grade K (*n* = 950)	Grade 1 (*n* = 1,821)	Grade 2 (*n* = 1,764)	Overall sample (*n* = 4,535)
Gender				
Male	131 (.14)	233 (.13)	255 (.14)	619 (.13)
Female	114 (.12)	224 (.12)	261 (.15)	629 (.14)
Unknown	705 (.74)	1,364 (.75)	1,248 (.71)	3,317 (.73)
Exceptionality				
SWD	11 (.01)	47 (.03)	42 (.02)	100 (.02)
Non-SWD	227 (.24)	395 (.22)	431 (.24)	1,053 (.23)
Gifted	7 (.01)	15 (.010	43 (.02)	65 (.01)
Unknown	705 (.74)	1,364 (.75)	1,248 (.71)	3,317 (.73)
Language				
ELL	22 (.02)	42 (.02)	20 (.01)	84 (.02)
Non-ELL	223 (.23)	415 (.23)	492 (.28)	1,130 (.25)
Unknown	705 (.74)	1,364 (.75)	1,252 (.71)	3.321 (.73)
Race				
White	70 (.07)	129 (.07)	172 (.10)	371 (.08)
Black	19 (.02)	31 (.02)	47 (.03)	97 (.02)
Asian	5 (.01)	3 (<.01)	3 (<.01)	11 (<.01)
Other	15 (.02)	30 (.01)	33 (.02)	78 (.02)
Unknown	841 (.89)	1,628 (.89)	1,509 (.86)	3,978 (.88)
Ethnicity				
Hispanic	29 (.03)	33 (.02)	19 (.01)	81 (.02)
Non-Hispanic	109 (.11)	193 (.11)	255 (.14)	557 (.12)
Unknown	812 (.85)	1,595 (.88)	1,490 (.84)	3,897 (.86)

Note. ELL = English language learner. SWD = Students with disabilities. Individual student demographic characteristics, such as ethnicity, exceptionality, or eligibility for free or reduced-price lunch, were not available for many students at the time of the present report was written. Some of the proportions do not sum to 1.00 because of rounding errors.

In the 2014–15 and 2015–16 school years, the Mathematics Florida Standards defined the official set of standards for mathematics in grades K–12 (Florida Department of Education, 2014). For the previous three school years, the CCSS-M (NGACBP & CCSSO, 2010) were the officially adopted curriculum standards for mathematics in Florida. The CCSS-M and the Mathematics Florida Standards are similar to one another but are not identical at these grade levels. No statewide assessment of student mathematics achievement in grades K–2 is conducted in Florida, but some individual districts use district-selected assessment tools to monitor progress of K–2 students.

2. Test Development, Scoring, and Data Entry Procedures

2.1. Content

The content standards at grades K, 1, and 2 in the CCSS-M (NGACBP & CCSSO, 2010) and Mathematics Florida Standards (Florida Department of Education, 2014) provide guidelines for content specifications. Overall, the focus of the test is on number, operations, and equality. It includes several items designed to favor students who have a solid grasp of place-value concepts. The highest number on the grade K test is 35, that on the grade 1 test is 102, and that on the grade 2 test is 686. Computation items presented symbolically involve applying the addition or the subtraction operation between 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, Fennema, Franke, Levi, & Empson, 1999; Carpenter et al., 2015).

Test design involved finding an optimal point at the intersection of two potentially competing goals: (1) to 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 K, 1 and 2 in the CCSS-M and the Mathematics Florida Standards and (2) to minimize the testing burden on teachers and students.

2.2. Instrument Development Process

The development process for the Spring 2016 K–2 EMSA tests consisted of the following phases:

1. Review of content expectations for grades K, 1, and 2 in the CCSS-M (NGACBP & CCSSO, 2010) and Mathematics Florida Standards (Florida Department of Education, 2014)
2. Review of the content and psychometric properties of the 2014 MPAC Interview (Schoen et al., 2016c), the 2015 MPAC Interview (Schoen et al., 2016c) and the Fall 2013, Fall 2014, and Fall 2015 EMSA test items (Schoen et al., 2016a, 2016b; Schoen et al., 2017)
3. Review of the content of the Cognitively Guided Instruction professional-development plan
4. Development of the first written draft of the test blueprint
5. Internal review of the draft blueprint by members of the evaluation team and external review by experts in mathematics and mathematics education
6. Revision of the blueprint based on feedback
7. Development of the first written draft of the test form for grades K, 1, and 2 and corresponding scoring procedures
8. Review of the draft test forms, editing, and proofing
9. Analysis of the frequency of correct response position and distribution of correct response positions across each grade-level test
10. Development of administration instructions
11. Proofreading of test and administration instruction forms

Test items from several tests previously administered in the fall or spring with grade 1 or 2 students informed the test in several ways. Items with poor psychometric statistics from previous field tests were not used. Many of the items on the previous field tests had an open-ended, constructed-response format. The students' responses in the open-ended format on previous test forms and pilot studies

informed the determination of the set of response options in the selected-response items of the Spring 2016 K–2 EMSA tests. In general, the five most frequently provided responses were used as the five response options on selected-response items.

During the process of expert review, test items were reviewed for content accuracy as well as potential 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 keywords (e.g., altogether, in all, left).

2.3. Test Design and Assembly

The Spring 2016 K–2 EMSA tests consisted of both selected-response and constructed-response items. The grade K test was composed of 14 selected-response and 7 constructed-response items. Of the 14 selected-response items, six of the items presented a set of five response options arranged horizontally across the page with exactly one correct response for each item. The other eight selected-response items presented two options: yes and no. The grade 1 test was composed of 16 selected-response and 10 constructed-response items. Eight of the selected-response items presented horizontally arranged sets of five response options, and the other eight presented the options yes and no. The grade 2 test was composed of 14 selected-response and 11 constructed-response items. Six of the items presented horizontally arranged sets of five response options, and the other eight presented the options yes and no.

The response options for selected-response items presenting five options were always numerals and were ordered left-to-right from least to greatest. The students were directed to fill in the circles (which were called "bubbles" in the administration guide) below their answer choices. Bubbles were centered beneath the corresponding response options, and responses were centered horizontally across the page.

A sample item with an example of responses is provided on the first page of the test for the administrator to use in demonstrating how students are expected to respond (e.g., by completely filling the bubble). The grade K test included an additional practice item as the first item on the test. This item asked students to fill in the bubble of the shape that was a triangle.

In general, the response options for the items in the Word Problems section include at least one of the two numbers given in the problem, their sum, and their difference. No pictures or images appear on the page apart from the page-numbering system, the text of the problem, the five numerals comprising the response options, the five ovals that students use to indicate their responses, and the boxes in which students record their responses on constructed-response items. Plenty of empty space is available on the page for students to draw or record their thoughts as necessary. Most of the items in the Computation and Number Facts sections and two of the items in the Operations on Both Sides of the Equals Sign section consist of items presented as open equations. Students were instructed to write their answer in the empty box representing the unknown value the equation to be solved.

In the Word Problems section, 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 visual information. For grades 1 and 2, in the Computation, Number Facts, and Equals Sign sections, several items are presented per page over multiple pages. The grammar used in word problems was reviewed by people with expertise in teaching emergent bilingual students. Futura font in large type size was used throughout. Copies of the grades K, 1, and 2 tests are presented in Appendices A, B, and C, respectively.

Pages were identified by a series of child-friendly images rather than page numbers. Figure 1 provides one example of these images. The large and easily distinguished image is also useful for the test administrator to use as a way to verify from across the room that students have turned to the correct page.

Figure 2.1. One of the images used in place of page numbers.

Test administrators were directed to read each math problem on the first part of the test aloud to students in accordance with the administration script. The second part of the test consisted of all constructed-response items, and students answered these questions at their own pace. In addition, administrators were asked to provide and allow students to use manipulatives, like counters or linking cubes, during the test. If students required testing accommodations resulting from IEP, ELL, or 504 plans, then the teacher was expected to provide any and all required accommodations for those individual students and to document the accommodation on the student information sheet. The test was not designed to be timed, so test administrators were instructed to allow students adequate time to answer all of the questions.

During the test development, careful consideration was given to the distribution of correct-response positions across each test form. Table 2.1 provides the number of times the correct answer is in each position at each grade level.

Table 2.1. Number of Times the Correct Answer Appears in Each Position

Grade level	A	B	C	D	E	Yes	No
K	0	1	1	4	2	6	2
1	0	3	2	1	2	6	2
2	0	4	1	1	3	6	2

2.4. Test Production

The tests, administration guides, and consent forms were printed on white 20-pound paper at Florida State University and distributed to the participating schools. Test forms for grades 1 and 2 were printed double-sided. Those for grade K were printed single-sided to reduce confusion among kindergarten students.

Test administration guides were created for each test and were grade-level-specific. The administration guide was repeatedly reviewed, edited, and proofread by research project staff during the test-development process.

2.5. Test Administration for the Spring 2016 EMSA K–2 Test

Each participating teacher was provided with a test packet containing

- Test-administration guide (for the corresponding grade level)
- Class set of student tests

- Student information sheet
- Parental consent forms (Districts A and B only)

These materials were distributed to the main offices at school sites during the first two weeks of April. They were distributed to the participating teachers through the main office personnel or principal-appointed designee. Teachers were instructed to administer the tests between April 18th and May 11th.

The Spring 2016 K–2 EMSA test administration guides provided an overview of the tests, described the administration process and directions, explained how to submit completed tests, and provided a full script to be read verbatim during administration of the test. The final forms of the test administration guides for grades K, 1, and 2 are presented in Appendices D, E, and F, respectively.

Student participation in the study was overseen by the human subjects committee (i.e., institutional review board) at Florida State University and by the school districts of participating students. Parental consent forms were distributed and collected in fall 2015. Some of the school districts approved or requested a passive-consent process. Two districts required and approved an active-consent process. In order to maximize the percentage of students in the participating classrooms who would be represented in the study, the test packets sent to classrooms in these two districts contained consent letters and forms to be distributed to those students who had not returned consent forms at the time of the Fall 2015 EMSA test administration.

Participating teachers received a class roster in the test packet. The roster indicated consent status of students in the class who were known to be permitted to participate in the study. Upon conclusion of test administration, teachers were instructed to submit all testing materials (test administration guide, used and unused student test booklets, student information sheet, and parental consent forms when applicable) to their principals or designees. Teachers were asked to return only test booklets completed by those students with parental consent. 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 third week of May 2016.

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. Two teachers were granted a time extension for materials pickup. The date of test administration was not used as a factor in data modeling.

2.6. Data Entry and Verification Procedures

Research assistants transferred student responses on the paper-based test forms into forms hosted on a FileMaker Pro database (FileMaker Pro, Version 14.1). Data for constructed-response items were manually keyed into the FileMaker forms as a numerical response. As a strategy to minimize data entry errors through typographical errors, response fields for selected-response items were validated to allow only the codes for offered responses, as well as codes for missing or uninterpretable responses.

Sometimes, students neglected to fill in the bubble corresponding to their responses. In these instances, research assistants examined written work provided by the student to determine whether a clear intended response seemed to have been given. When a clearly intended response matched one of the response options given, it was then coded as though the student had filled in the bubble. For example, when an examinee wrote a numeral in the position of the unknown variable in line with the equation and did not fill the bubble below the corresponding numeral, the response was entered as though the bubble had been filled. When the intended response did not match one of the response options

presented, the response was coded as unclear intent (UI). If no intended response was evident, the response was coded as did not respond (DNS). The code MR, for multiple response, was also used when students selected more than one response from the options presented and did not clearly attempt to erase one of the responses or cross out their unintended responses.

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. Most of the students' responses to the constructed-response items were entered as the student wrote them, but research assistants could choose one of the missing-item codes, UI and DNS, when applicable.

To verify the accuracy of the entered data, research assistants entered the results from 11% of the tests twice. These tests were selected through a random sample of participating classes, stratified by grade level. Two different assistants entered data for the first and second entries. This procedure ensured that the double-entry comparison was performed on a sample representing all three grade levels and all the data-entry personnel. These separate records were compared and found to have an overall 99.0% agreement on items entered.

2.7. Item-scoring Procedures

Responses to selected-response items were scored according to the scoring guide provided in Appendix G of this report. Student responses to each constructed-response item were aggregated and reviewed by an adjudication committee composed of experts in mathematics to identify any unanticipated—but mathematically correct—responses.

More than 99.2% of the test forms in the sample had responses for more than three-quarters of the items. The no-response rate was highest in Kindergarten, where less than 98.3 percent of examinees provided a response for more than three-quarters of the items. Item responses coded as DNS, UI, and MR were consequently scored as incorrect, and all available data were used in the subsequent analyses.

3. Data Analytic Procedures

After the test data were entered, scored at the item level, and verified for accuracy, the data from the field test of the Spring 2016 EMSA were subjected to the following analyses:

1. Initial screening of items with classical test theory (CTT)
2. Exploratory Factor Analysis (using tetrachoric correlations to preclude difficulty-related factors)
3. Within-grade scaling with a two-parameter logistic item-response theory (2PL-IRT) model
4. Equating of scales between grades by means of a nonequivalent groups with anchor tests (NEAT) design (i.e., common items between grades) to create the vertical scale. The Stocking-Lord method (Kolen & Brennan, 2014) was used to transform the within-grade scales to the common, vertical scale.

Initial item screening with CTT was completed to identify items that might not be providing useful information about test-takers' abilities (e.g., overly difficult or easy items). Factor analysis tested the dimensionality of the test as a means of determining whether the test was measuring a sufficiently unidimensional construct (see Anderson, Kahn, & Tindal, 2017). This analysis informed whether we would generate scale scores for a unidimensional or a multidimensional construct. As described in greater detail below, the results of the factor analyses supported an essentially unidimensional measure, and scaling proceeded accordingly.

All analyses and displays of data were conducted within the R statistical computing environment (R Core Team, 2017). The following sections provide more detailed information about the analytic processes we used in each phase of analysis.

3.1. Initial Screening by Means of Classical Item Analysis

Using an approach based on CTT, we generated several statistics for each item on the basis of the sample for each separate grade level. These statistics provided empirical information about the quality of each item. As described in the subsequent sections, we set thresholds (i.e., p-value < .10, p-value > .90, point estimate for point-biserial correlation < .20) to determine which items to consider for deletion on the basis of the results. These thresholds did not establish bright-line rules for inclusion or exclusion. Rather, items that were close to these thresholds were marked for further analysis and discussed by the development team. The item statistics and the relation between the item and the test as a whole were considered with respect to whether an item remained or was removed.

3.1.1. Classical item difficulty

Each individual item on the Spring 2016 K–2 EMSA was scored dichotomously. For these items, the CTT-based item difficulty statistic, or p-value, corresponds to the proportion of test takers in the within-grade-level samples who produced a correct answer to the item. Desirable p-values typically fall between .10 and .90, but these boundaries serve as guidelines rather than strict rules. Items with particularly high or low p-values may not be contributing useful information to the overall score, but that is not always the case. At times, those high- or low-difficulty items may be useful for discriminating among test-takers in the corresponding ability range (i.e., very high or low achievement levels). Items scoring below/above these thresholds were more closely examined.

3.1.2. Classical item discrimination

Items are considered to have good discrimination if high-ability students tend to answer correctly and low-ability students tend to answer incorrectly. According to a classical approach, the item discrimination was assessed by examination of the relation between test-takers' performance on each individual item and their total raw scores (total number of correct items). This correlation was calculated for each item on each test. The point-biserial correlation is interpreted similarly to any other correlation; values fall between negative one and positive one. Generally, point-biserial correlations are positive, indicating that students with a higher score (i.e., higher ability) are more likely to respond to the item correctly. Items with negative point-biserial correlations are highly concerning, because they indicate exactly the opposite—as students ability increases, their likelihood of responding correctly to the individual item decreases. In practice, negative values are rare, but any value less than .20 is cause for concern. All items with point-biserial correlations less than (or near) .20 were marked for review during the item screening process.

3.1.3. Item/raw score plots

Additional screening involved the generation of item/raw-score plots, where students' total scores were plotted along the horizontal axis, and the proportion responding correctly was mapped onto the vertical axis. Separate lines were produced for each item. (See Appendix H.) Because the sample size for each individual raw score was relatively low, we smoothed the overall relation using local scatterplot smoothing (loess), such that the overall trend could be examined. Items with shallow, negative, or u-shaped slopes were identified and further scrutinized.

3.2. Exploratory Factor Analysis

The primary goal of the analyses reported here was to create a unified, vertical scale spanning grades K–2, such that scores on the grade K, grade 1, and grade 2 tests would be directly comparable. We constructed this scale using IRT, as described below. One of the primary assumptions of IRT, however, is local independence of item responses, implying that students' probability of success on any one item is independent of their probability of success on any other items on the test, conditional on ability. Local dependence can inflate construct-irrelevant variance and reliability estimates. When a standard unidimensional model is fit—as was the goal here—extra dimensions in the data can lead to local item dependence and threaten the stability of the scale. As a preliminary step, before creating the vertical scale, we explored the dimensionality of each scale.

Because all items were dichotomous, tetrachoric correlation matrices were used to help protect against arriving upon difficulty-related factors rather than substantive factors. When evaluating how many factors to retain, we compared three tests: Velicer's minimum average partial test (MAP; Velicer, 1976), Revelle's very simple structure test (VSS; Revelle & Rocklin, 1979), and parallel analysis (Horn, 1965). In cases where these three tests provided conflicting evidence in terms of the optimal number of factors to extract, scree tests were used as an arbiter. All models were fit with maximum likelihood by means of an oblique rotation (implying that, when multiple factors were extracted, they were allowed to be correlated). Models were estimated within the R statistical environment (R Core Team, 2017) by means of the *psych* package (Revelle, 2017). Results of these analyses are presented in Table 3.1 and Figures 3.1, 3.2, and 3.3.

Table 3.1. Number of Factors Suggested by the MAP and VSS Tests

Grade level	MAP	VSS1	VSS2
K	2	2	2
1	2	1	2
2	2	1	2
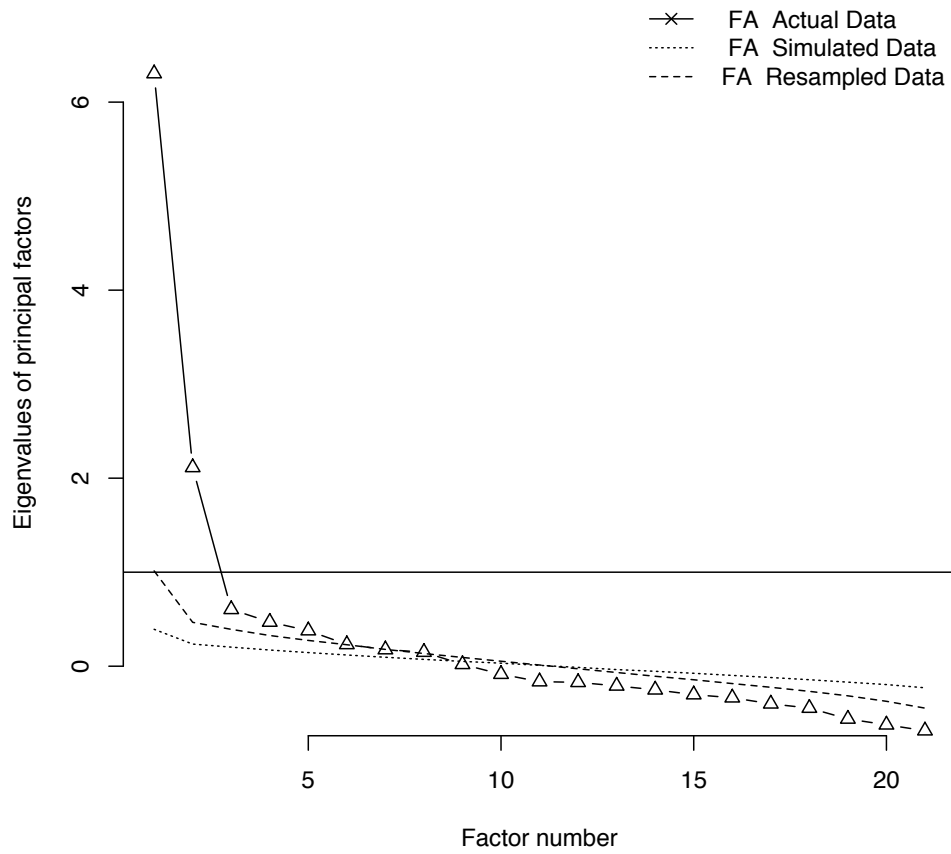
Figure 3.1. *Parallel analysis scree plot for the grade K test.*
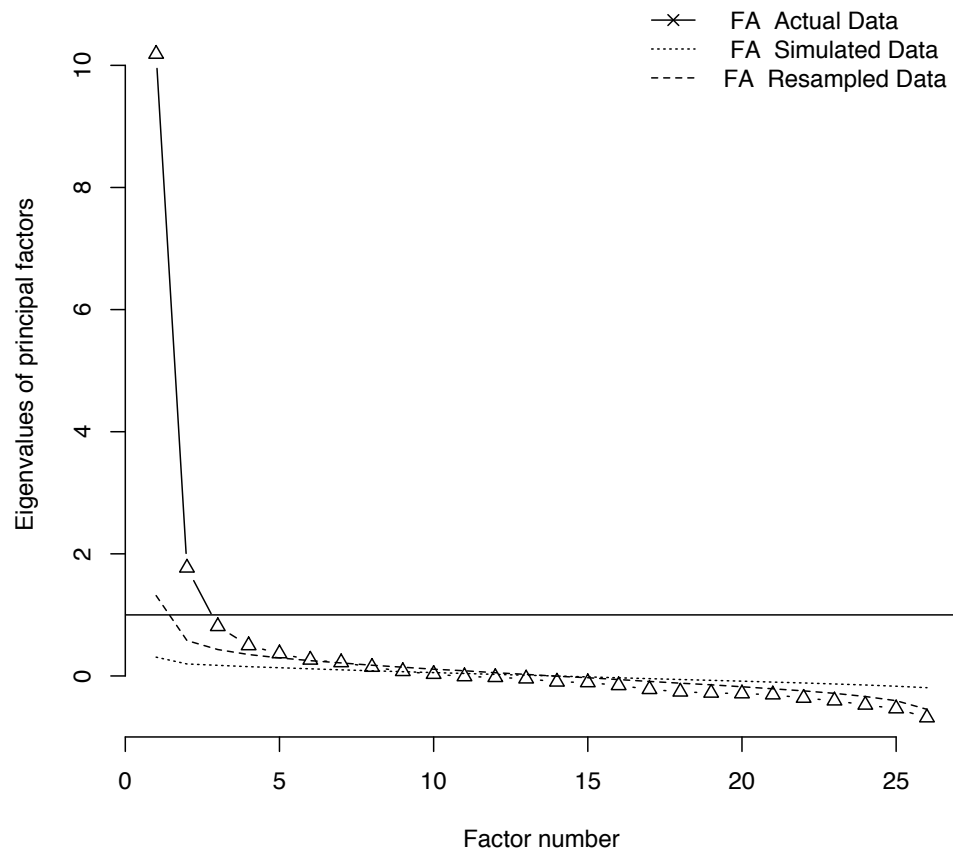
Figure 3.2. Parallel analysis scree plot for the grade 1 test.
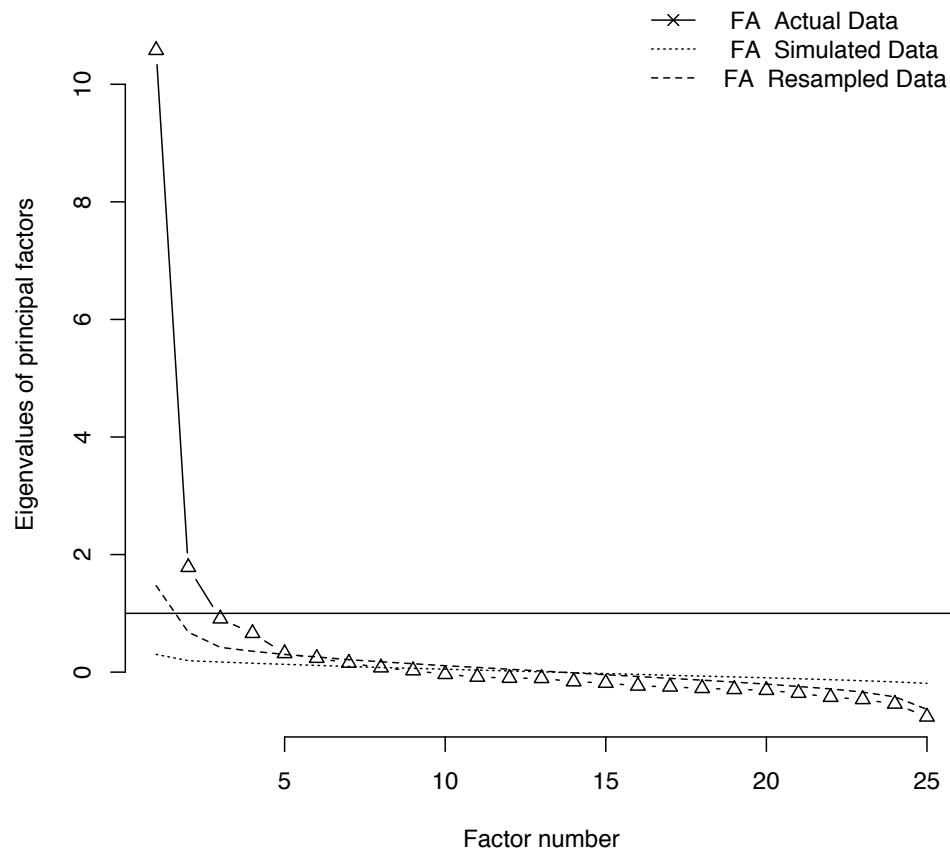
Figure 3.3. Parallel analysis scree plot for the grade 2 test.

Across all tests of the number of dimensions to extract, one or two dimensions were estimated as optimal. The scree plots all displayed a large drop in the eigenvalues after extraction of the first dimension, although the eigenvalue from the second dimension extracted was universally greater than the eigenvalue from the second dimension of the randomly generated data (i.e., parallel analysis always indicated more than one dimension). Collectively, these results indicated that the test was likely to be uni- or bidimensional. Recent evidence from Anderson et al. (2017), however, suggests that the 2PL IRT model is robust to mild deviations from unidimensionality. Given that the purpose of the scaling was to create a single scale across grades K–2, we proceeded to IRT scaling by assuming a unidimensional structure, although we recognized that a two-dimensional structure might have better represented the underlying data in some cases (e.g., Grade K, where all tests indicated two dimensions).

When the factor loadings relative to the items were inspected, a relatively consistent "item format" dimension seemed to emerge; standard multiple-choice items loaded on one dimension and "fill in the blank" items on a second. One method of handling these would be to model additional "item format" dimensions (one dimension for each format) with all items still loading on a single dimension to yield the vertical scale. This method is essentially equivalent to that employed by Anderson et al. (2017), who

found that, even though the model fit the data better, the two models were nearly indistinguishable in both item and person parameter estimates. We therefore proceeded with the unidimensional model, assuming that the 2PL model was robust to any potential violations of local item dependence that may have arisen as a result of the item-formatting issue.

3.3. Specification of Models Based on Item-response Theory

After the exploratory factor analyses, we fit a unidimensional 2PL-IRT model to the data within each grade separately. The basic model was fit in accordance with Equation 1,

$$P(y_{ij} = 1 | \theta_j, a_i, b_i,) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \tag{1}$$

where θ_j represents the estimated ability of student j, and a_i and b_i are the discrimination and difficulty of item i, respectively. In essence, the log odds of students' correctly responding to an item are driven by the difference between their estimated ability, θ_j, and the difficulty of the item b_i. Log odds are estimated as the ratio between the odds of a correct and those of an incorrect response. The discrimination parameter represents the slope of the item characteristic curve (i.e., the rate at which the probability of a correct response changes as θ increases). Items with lower discrimination values are weighted less in the estimation of θ than those with higher values, as the difference between the item difficulty and the students' ability is multiplied by the estimated discrimination of the item. Generally, the discrimination index is viewed as an "item quality" index; higher discrimination (to a point) indicates higher quality and is therefore weighted more heavily in the estimation of students' ability.

After the initial 2PL model was fit, a few items had overly high discrimination values with broad standard errors (e.g., 3.61 with a standard error of 0.23). The values provided evidence that the model was potentially overly complex for the data. Further examination of these items revealed that, in the majority of cases, items with imprecise discrimination estimates were those related to the equals sign. We therefore simplified the models by constraining the discrimination parameter among these items to be equal, which resulted in much lower estimated standard errors and more reasonable discrimination estimates. In other words, the final model fit was a modified 2PL model, with the discrimination parameter freely estimated for the majority of items, but constrained to be equivalent for all items relating to the equals sign.

3.4. Vertical Linking

After arriving at a final scale for each grade, we equated the scales to establish the vertical scale using the items common to different grades. We centered the scale on grade 1—the middle of the grade span—and equated both the grade K and grade 2 test parameters relative to the grade 1 scale. Because all grade-level test forms included common items, multiple links joined each test and the grade 1 scale. That is, the grade K test included a direct link of common items between grades K and 1, but also an indirect link through the common items with grade 2. Similarly, grade 2 included both a direct and an indirect link with grade 1. Rather than using just the direct links, we used a weighted combination of the two, weighting them by the standard error of the equating coefficient. This method, known as the weighted bisector method, can lead to more accurate estimates by using all the information in the data, rather than just the information provided by the direct links (see Battauz, 2013). In our specific case, however, because only one direct and one indirect link were available, and the indirect link was associated with a higher standard error (and thus weighted less), the difference between using both links and using just the direct link was almost indistinguishable.

Equating coefficients were estimated by the Stocking-Lord method, which uses the test characteristic curves to derive the coefficients. These coefficients were used to transform item and person parameters in grades K and 2 onto the grade 1 scale by means of standard transformation procedures (see Kolen & Brennan, 2014).

4. Results

4.1. Initial Screening of Items

The first step in data analysis involved reviewing the proportion correct and point-biserial statistics for each item on the grade K, 1, and 2 tests. These statistics were based on the within-grade samples for their corresponding grade levels. This initial screening process revealed a fairly even spread of item difficulties (as defined by percentage correct within the sample), including some items answered correctly by almost all of the respondents and some answered correctly by very few. These statistics are given in Appendix H for all items on the test.

For brevity, we discuss only those items removed from the scales during the screening process. Those items, along with their p-values and point-biserial statistics, are listed in Table 4.1. Six of the items—the same two items on each of the three test forms—that were removed from the scales were intended to serve as practice items for students to learn how to respond to the equals-sign questions. Two of the items that were removed from individual grade level scales had been intended to remain in the final scale in the original design of the test. G1i5_G2i3 was removed from the final scale for the grade 1 test, and GKi22_G1i25_G2i23 was removed from the final scale for the grade K test. G1i5_G2i3 was removed from the grade 1 scale, because the trace line in the spaghetti plot was u-shaped. (See Figure H.2 in Appendix H.)

Table 4.1. Classical Test Theory–Based Item Statistics for Items Removed from Scale during Screening Process

Item	Item description	Grade level	p (se)	PB
GKG1G2i6		K	.92 (.009)	.31
GKG1G2i7		K	.90 (.010)	.31
GKi22_G1i25_G2i23		K	.14 (.011)	.39
G1i5_G2i3		1	.48 (.012)	.54
GKG1G2i6		1	.99 (.003)	.17
GKG1G2i7		1	.99 (.003)	.19
GKG1G2i6		2	.99 (.002)	.14
GKG1G2i7		2	.99 (.002)	.15

Table 4.2 shows the distribution of item difficulty and item discrimination, as measured by the point-biserial correlation, for the items used at each of the three grade levels after items were culled based on the initial screening process. Figures 4.1, 4.2, and 4.3 show the raw-score distributions for students on the total test score for grades K, 1, and 2, respectively.

Table 4.2. Distribution of Item Difficulties and Discrimination Point Estimates for Items Used in the Final Scales

Bins	Number of items		
	Grade K	Grade 1	Grade 2
P-value			
>.90	1	2	3
.80–.89	2	5	5
.70–.79	1	7	5
.60–.69	2	4	3
.50–.59	3	2	4
.40–.49	6	1	1
.30–.39	2	1	2
.20–.29	1	1	0
.10–.19	0	0	0
<.09	0	0	0
Mean	0.57	0.69	0.70
Median	0.50	0.74	0.74
SD	0.22	0.17	0.18
Point-biserial correlation			
.80–1.0	0	0	0
.60–.79	0	6	5
.40–.59	13	13	15
.20–.39	5	4	3
.00–.19	0	0	0
Mean	0.45	0.50	0.49
Median	0.44	0.51	0.51
SD	0.09	0.10	0.15

Note. P-value = proportion of sample judged to have provided a correct answer.
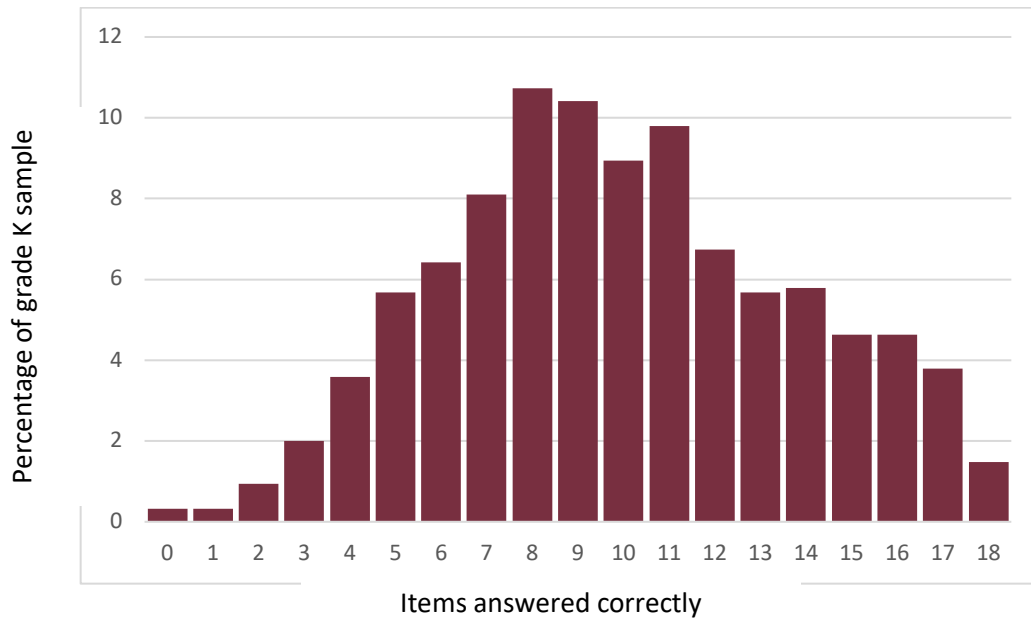
Figure 4.1. Distribution of the number of items answered correctly in the final, 18-item scale administered to the grade K sample (n = 950).
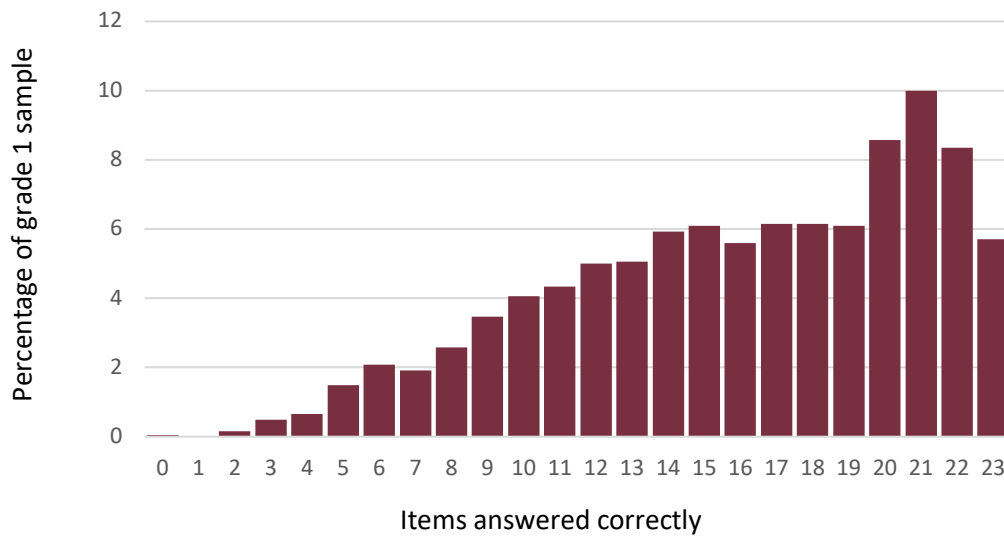
Figure 4.2. Distribution of the number of items answered correctly in the final, 23-item scale administered to the grade 1 sample (n = 1,821).
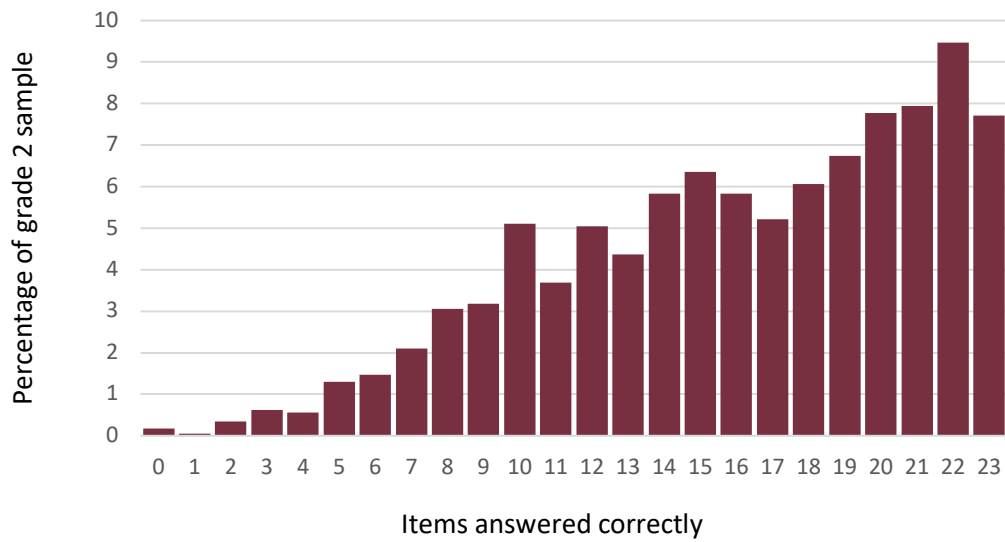
Figure 4.3. Distribution of the number of items answered correctly in the final, 23-item scale administered to the grade 2 sample (n = 1,764).

4.2. Item Response Theory Models

The IRT-based discrimination and difficulty estimates for grades K, 1, and 2 are presented in Tables 4.3, 4.4, and 4.5, respectively. The mean item difficulties based on the within-grade models were -1.13, -0.90, and -0.46 for grades K, 1, and 2, respectively. Item discriminations ranged from 0.94 to 4.91 in grade K, 0.86 to 1.93 in grade 1, and 0.84 to 2.18 in grade 2.

Table 4.3. Grade K Vertical and Within-Grade Scales IRT Estimates

Item	Vertical scale		Within-grade level scale	
	Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
GKi2	3.01 (.173)	−1.84 (.171)	1.53 (.173)	−2.52 (.171)
GKi3_G1i1	3.20 (.170)	−1.67 (.153)	1.63 (.170)	−2.13 (.153)
GKi4_G1i2	2.61 (.125)	−.90 (.088)	1.33 (.125)	.27 (.088)
GKi5_G1i3_G2i1	2.25 (.123)	−.34 (.106)	1.15 (.123)	1.48 (.106)
GKG1G2i8	.94 (.038)*	−.75 (.069)	.48 (.038)	.24 (.069)
GKG1G2i9	.94 (.038)*	−.94 (.068)	.48 (.038)	.06 (.068)
GKG1G2i10	.94 (.038)*	−1.98 (.075)	.48 (.038)	−.92 (.075)
GKG1G2i11	.94 (.038)*	−1.00 (.068)	.48 (.038)	<.01 (.068)
GKG1G2i12	.94 (.038)*	−1.50 (.070)	.48 (.038)	−.47 (.070)
GKG1G2i13	.94 (.038)*	−.49 (.070)	.48 (.038)	.48 (.070)
GKG1i14	2.82 (.132)	−.90 (.091)	1.44 (.132)	.30 (.091)
GKG1i15	1.73 (.097)	−.83 (.077)	.88 (.097)	.30 (.077)
GKi16	3.99 (.241)	−1.85 (.265)	2.03 (.241)	−3.39 (.265)
GKG1i17	4.91 (.250)	−1.30 (.159)	2.50 (.250)	−1.48 (.159)
GKG1i18_G2i15	3.62 (.170)	−1.13 (.105)	1.84 (.170)	−.46 (.105)
GKi19	2.71 (.129)	−1.17 (.091)	1.38 (.129)	−.44 (.091)
GKi20_G1i19_G2i16	2.81 (.133)	−.91 (.090)	1.43 (.133)	.25 (.090)
GKi21_G1i22	3.59 (.170)	−.80 (.108)	1.83 (.170)	.74 (.108)

*Discrimination parameters for these items were fixed to .94 during the IRT model calibration process. See section 3.3 for explanation.

Table 4.4. Grade 1 Vertical and Within-Grade Scales IRT Estimates

Item	Vertical scale		Within-grade level scale	
	Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
GKi3_G1i1	1.25 (.170)	−2.99 (.195)	1.25 (.170)	−3.74 (.195)
GKi4_G1i2	1.90 (.115)	−.58 (.090)	1.90 (.115)	−1.10 (.090)
GKi5_G1i3_G2i1	1.76 (.105)	−.46 (.079)	1.76 (.105)	−.81 (.079)
G1i4_G2i2	1.62 (.098)	−.48 (.075)	1.62 (.098)	−.79 (.075)
GKG1G2i8	1.73 (.051)*	−.54 (.074)	1.73 (.051)*	−.94 (.074)
GKG1G2i9	1.73 (.051)*	−1.12 (.082)	1.73 (.051)*	−1.93 (.082)
GKG1G2i10	1.73 (.051)*	−1.34 (.088)	1.73 (.051)*	−2.32 (.088)
GKG1G2i11	1.73 (.051)*	−.91 (.078)	1.73 (.051)*	−1.57 (.078)
GKG1G2i12	1.73 (.051)*	−1.26 (.085)	1.73 (.051)*	−2.18 (.085)
GKG1G2i13	1.73 (.051)*	−.28 (.072)	1.73 (.051)*	−.48 (.072)
GKG1i14	1.93 (.126)	−.94 (.108)	1.93 (.126)	−1.82 (.108)
GKG1i15	1.58 (.107)	−1.08 (.094)	1.58 (.107)	−1.70 (.094)
G1i16_G2i14	1.73 (.109)	−.80 (.089)	1.73 (.109)	−1.37 (.089)
GKG1i17	1.49 (.154)	−2.26 (.172)	1.49 (.154)	−3.37 (.172)
GKG1i18_G2i15	.93 (.092)	−2.19 (.088)	.93 (.092)	−2.04(.088)
GKi20_G1i19_G2i16	1.29 (.098)	−1.42 (.090)	1.29 (.098)	−1.82 (.090)
G1i20_G2i17	1.01 (.079)	−1.23 (.069)	1.01 (.079)	−1.24 (.069)
G1i21_G2i18	1.32 (.095)	−1.22 (.085)	1.32 (.095)	−1.61 (.085)
GKi21_G1i22	1.56 (.112)	−1.29 (.104)	1.56 (.112)	−2.02 (.104)
G1i23_G2i19	1.12 (.074)	−.20 (.059)	1.12 (.074)	−.22 (.059)
G1i24_G2i20	.86 (.069)	1.23 (.062)	.86 (.069)	1.06 (.062)
GKi22_G1i25_G2i23	1.73 (.051)*	.33 (.073)	1.73 (.051)	.58 (.073)
G1i26_G2i24	1.73 (.051)*	.42 (.073)	1.73 (.051)	.72 (.073)

*Discrimination parameters for these items were fixed to 1.73 during the IRT model calibration process. See section 3.3 for explanation.

Table 4.5. Grade 2 Vertical and Within-Grade Scales IRT Estimates

Item	Vertical scale		Within-grade level scale	
	Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
GKi5_G1i3_G2i1	1.45 (.114)	−.94 (.107)	1.53 (.114)	−2.11 (.107)
G1i4_G2i2	1.33 (.096)	−.57 (.083)	1.39 (.096)	−1.43 (.083)
G1i5_G2i3	1.38 (.089)	.06 (.070)	1.45 (.089)	−.61 (.070)
G2i4	1.61 (.118)	−.70 (.107)	1.70 (.118)	−1.95 (.107)
G2i5	1.34 (.086)	.24 (.067)	1.41 (.086)	−.36 (.067)
GKG1G2i8	2.18 (.066)*	−.08 (.090)	2.29 (.066)	−1.28 (.090)
GKG1G2i9	2.18 (.066)*	−.89 (.107)	2.29 (.066)	−3.05 (.107)
GKG1G2i10	2.18 (.066)*	−.59 (.099)	2.29 (.066)	−2.39 (.099)
GKG1G2i11	2.18 (.066)*	−.54 (.098)	2.29 (.066)	−2.29 (.098)
GKG1G2i12	2.18 (.066)*	−.93 (.108)	2.29 (.066)	−3.12 (.108)
GKG1G2i13	2.18 (.066)*	.04 (.088)	2.29 (.066)	−1.00 (.088)
G1i16_G2i14	1.75 (.137)	−.92 (.133)	1.84 (.137)	−2.50 (.133)
GKG1i18_G2i15	.85 (.117)	−2.75 (.119)	.89 (.117)	−2.77 (.119)
GKi20_G1i19_G2i16	1.17 (.127)	−1.92 (.133)	1.23 (.127)	−2.85 (.133)
G1i20_G2i17	.95 (.095)	−1.64 (.091)	1.00 (.095)	−2.04 (.091)
G1i21_G2i18	1.59 (.147)	−1.40 (.157)	1.68 (.147)	−3.04 (.157)
G1i23_G2i19	1.18 (.087)	−.52 (.075)	1.24 (.087)	−1.21 (.075)
G1i24_G2i20	.99 (.072)	.17 (.059)	1.04 (.072)	−.33 (.059)
G2i21	1.37 (.089)	1.21 (.073)	1.44 (.089)	.96 (.073)
G2i22	.84 (.066)	.28 (.056)	.88 (.066)	−.19 (.056)
GKi22_G1i25_G2i23	2.18 (.066)*	.45 (.086)	2.29 (.066)	−.13 (.086)
G1i26_G2i24	2.18 (.066)*	.50 (.086)	2.29 (.066)	−.02 (.086)
G2i25	2.18 (.066)*	.93 (.088)	2.29 (.066)	.93 (.088)

*Discrimination parameters for these items were fixed to 2.18 during the IRT model calibration process. See section 3.3 for explanation. The astute reader will notice that G2i22 was an equals-sign item, and G2i25 is not, but the discrimination parameter for G2i25 is fixed to the same value as those of the equals-sign items, but the discrimination parameter for item G2i22 is not fixed to the same value as the equals-sign items. This discrepancy was the result of a clerical error that was caught too late for the models to be rerun. The parameter estimates are reported the way the models were run.

After the within-grade scaling, equating coefficients to transform each of the grade K and grade 2 scales to the grade 1 scale were estimated, by means of the Stocking-Lord method with the weighted bisector approach, as described previously. The A and B coefficients are reported in Table 4.6, along with their standard errors. These coefficients were used to transform the within-grade scales to a common, vertical scale that is directly comparable across grades.

Table 4.6. Scaling Coefficients Used to Transform the Within-Grade Scales to a Common, Vertical Scale

From	To	A (SE)	B (SE)
K	1	0.57 (0.03)	−1.11 (0.04)
2	1	1.14 (0.04)	0.56 (0.04)

Figure 4.4 displays the test characteristic curves for each of the three grade levels on the vertical scale. Dashed vertical reference lines represent the inflection points on the scale (i.e., the ability level at which

students would be expected to get more than half the items correct). We note that the number of items differed for different grade levels, affecting the heights of the curves in Figure 4.4 (highest possible score). The curves indicate separation between grade levels, a desirable feature of the scale, especially at these grade levels (where students change and learn very quickly). The vertical lines can be interpreted as the estimated ability level associated with having a 50% chance of responding to approximately half the items correctly.
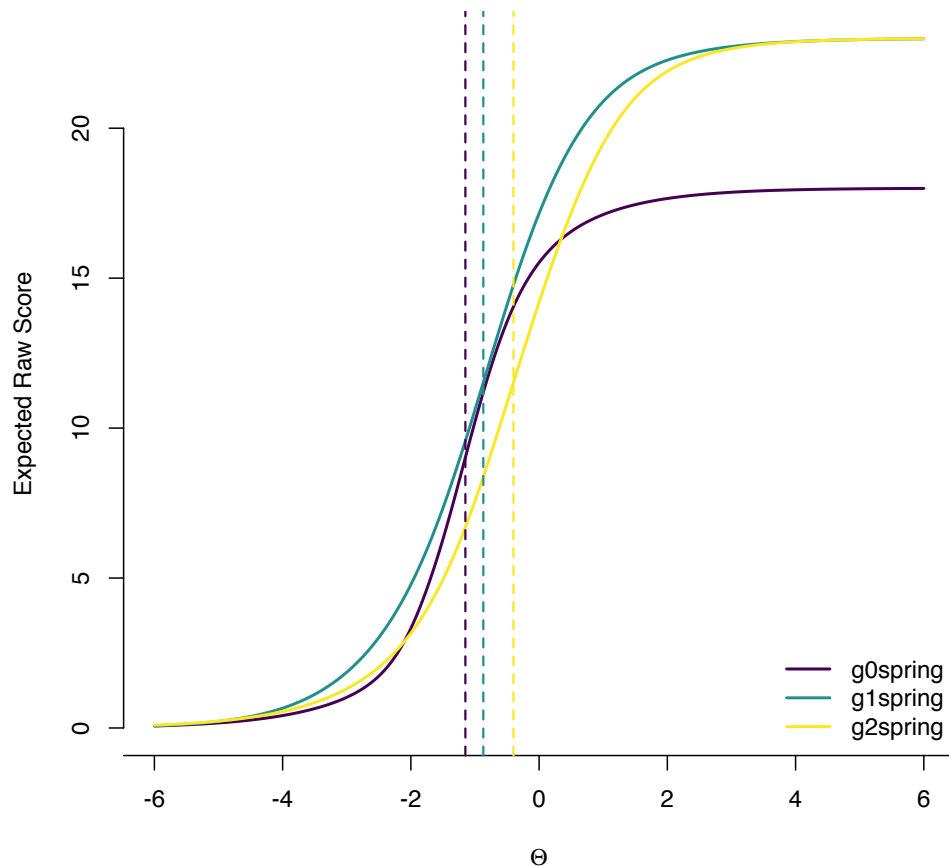
Figure 4.4. Test characteristic curves for grades K, 1, and 2.

4.3. Reliability

Item response theory provides a conditional view of reliability, in which the reliability of the measure is viewed as depending on the ability level of the respondent. This approach recognizes that reliability is not fixed but variable; it depend on who is taking the test. Figure 4.5 below displays the test information functions for each of the three tests. These functions are test-level summaries of the reliability, each mapped on the common, vertical scale. Under the standardized θ, reliability is equivalent to a Cronbach's alpha of 0.80 when information is equal to 5.0. Therefore, in the figure below, vertical

dashed lines display the ability regions for each test in which information is greater than or equal to 5.0 (implying the ranges in which reliability is ≥ 0.80).
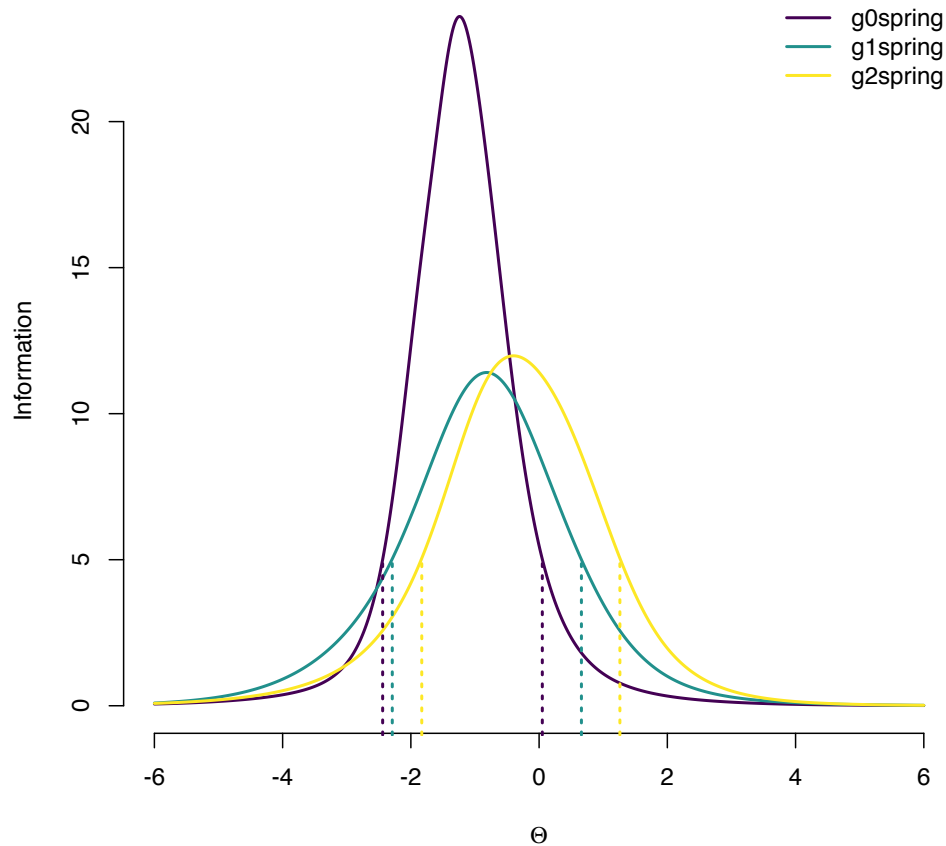
Figure 4.5. Test information functions for Grades K, 1, and 2.

Another way to understand the reliability of the instrument is to examine the match between the distributions of person abilities and item difficulties. In Figures 4.6, 4.7, and 4.8, the distributions for each are displayed on the common scale. At grade K, the distribution of items matched the distribution of persons quite well, so the vast majority of the person distribution was within the reliable range. At grades 1 and 2, the person distribution was shifted slightly to the right of the item distribution, implying students generally had a higher ability than the estimated difficulty of the items. The upper portion of the person distribution was therefore outside of the reliability range. Across all plots, the ranges in which reliability was equivalent to 0.80 or above are displayed. The vertical yellow lines in Figures 4.6 to 4.8 display this range. Students lying outside of these ranges (to the left of the left line and to the right of the right line) represent students for whom the test was equivalent to less than 0.80 reliability.
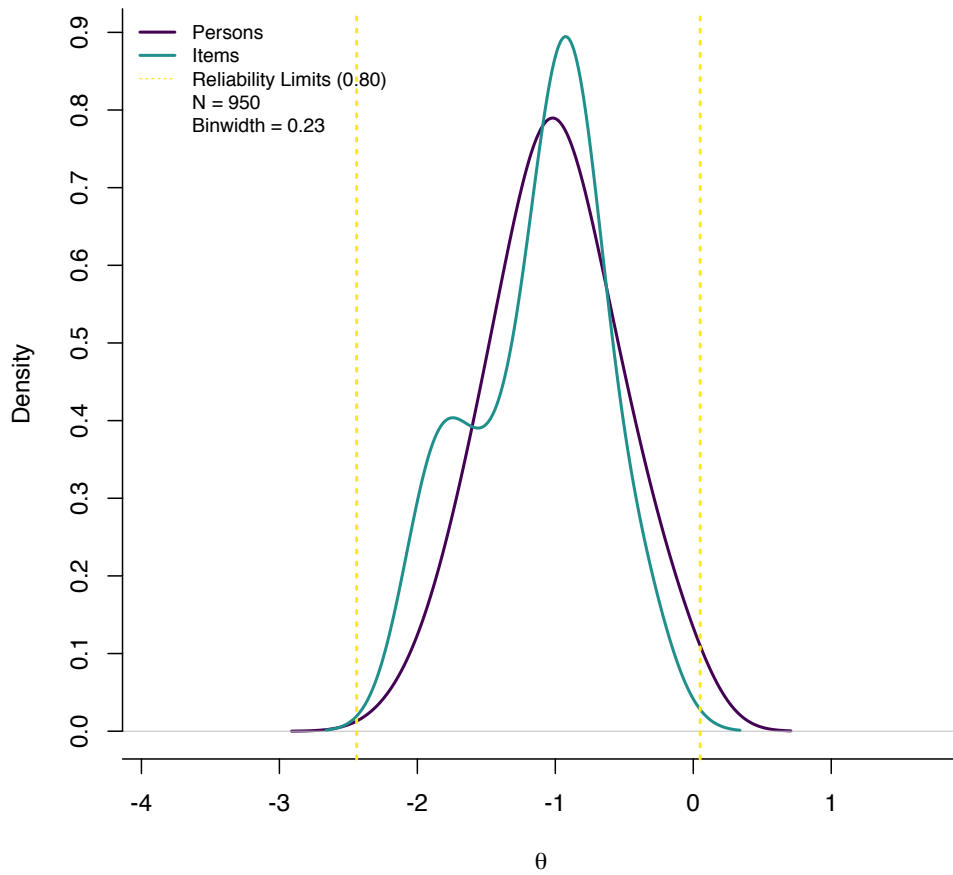
Figure 4.6. Grade K item-person plot.
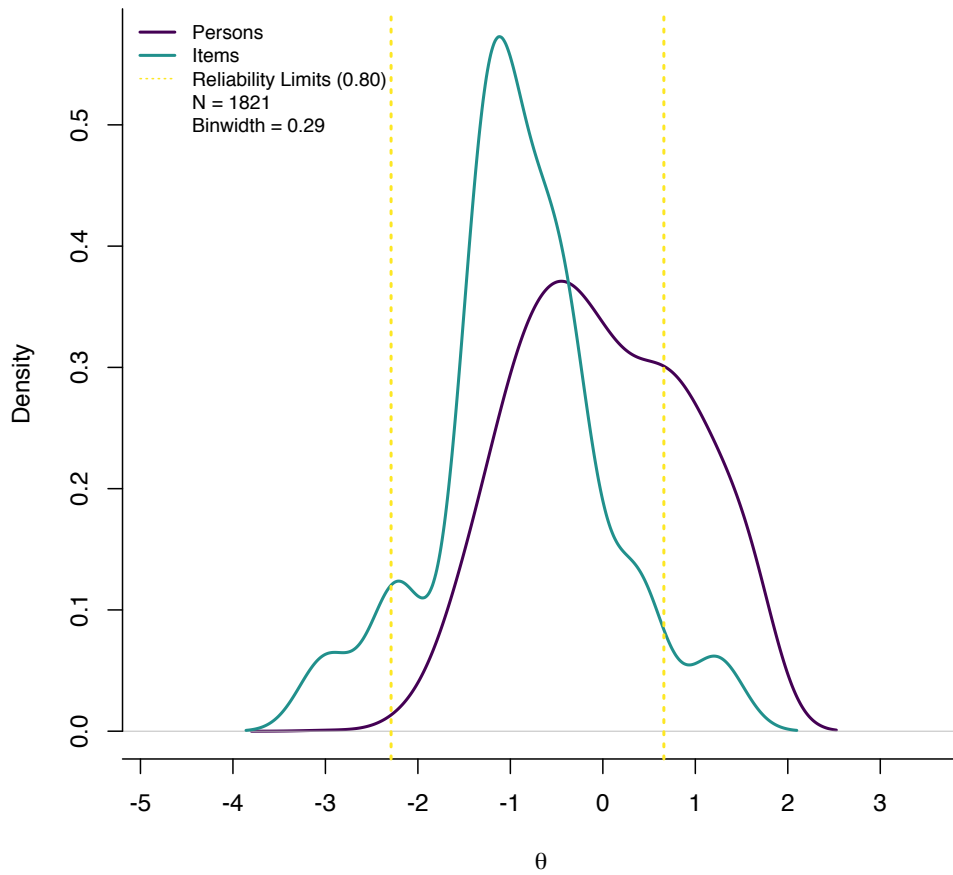
Figure 4.7. Grade 1 item-person plot.
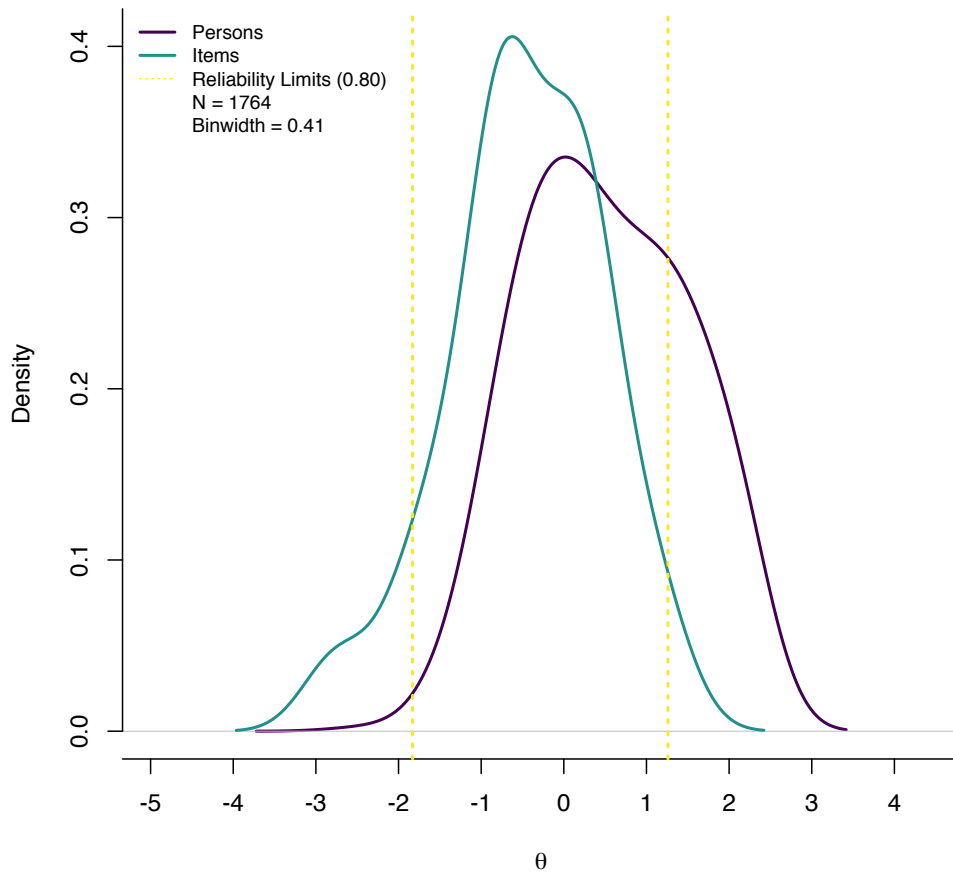
Figure 4.8. Grade 2 item-person plot.

5. Discussion and Reflection

The Spring 2016 K–2 EMSA was designed to be used as an end-of-year student achievement test for students in grades K, 1, and 2 in the area of number, operations, and algebraic thinking. Feasibility tests indicate the Spring 2016 K–2 EMSA can be successfully implemented in typical classrooms at a large scale.

The content of the test is mostly aligned with expectations of students in those grade levels and domains in accordance with the Common Core State Standards for Mathematics (NGACBP & CCSSO, 2010). In a few instances, the content of the test extends beyond the content specified by the CCSS-M for the given grade level. For example, there are word problems at each grade level involving *multiplication-grouping* or *measurement division* situations (or both) (Carpenter et al., 2015). Although this problem type is not specifically referred to in the CCSS-M at those respective grade levels, students do have the ability to solve them, and the items generally aim to measure student ability to solve these types of problems in situations involving grouping by tens, which is a key component of place-value understanding—a major topic in the grades K–2 CCSS-M. The tests also includes items designed to measure student understanding of the equals sign in mathematics—a topic that is not explicitly included in grade K. The equals-sign items go slightly beyond a strict interpretation of the specified content in grade 1 of the CCSS-M, because they involve questions about four numerical quantities rather than just three. These items were included on the test, because they are relevant to topics at these grade levels, and many students at those grade levels are able to solve them.

Tests of dimensionality show that the Spring 2016 K–2 EMSA test may be bidimensional, but follow-up investigations revealed that the multidimensionality was probably an item-formatting (i.e., selected-response, constructed-response) effect, which the 2PL is generally robust to.

Development of the Spring 2016 K–2 EMSA was based on previous tests of student ability in grades K, 1, and 2 (Schoen et al, 2016b; 2016d; 2017). The Spring 2014 and Spring 2015 MPAC tests were not equated across grade levels, so the Spring 2016 K–2 EMSA represents the first test in this series that was designed to be vertically equated and scored on a single scale that spans the grade levels. The equating process was successful, and the common scale affords considerable support for the use of the EMSA tests as a student achievement test in randomized trials spanning multiple grade levels.

The difficulty of the grade K test appeared to match the ability of the sample of grade K students well. This result can be seen in most clearly in Figure 4.6 but it also in the distributions of item difficulties in Table 4.2 and Figure 4.2. The difficulty level of the grade-1 and -2 tests appear to be lower than the ability level of the grade 1 and 2 students who took the test, as can be seen in Figures 4.3, 4.4, 4.7, and 4.8. More than 5% of grade 1 and grade 2 students answered every question correctly, suggesting a ceiling on the test and introducing concerns about reliability of the estimates of student abilities in the upper range of the distribution. At grade 1, approximately 27% ($n = 497$) of students scored above the 0.80 threshold for reliability; only 1 student scored below the 0.80 reliability range. A similar pattern was evident at grade 2; approximately 27% ($n = 475$) of grade-2 students in the sample scored above the 0.80 threshold, and only 5 below the 0.80 threshold. In contrast all students in grade K scored within the 0.80 threshold.

Further development to align the difficulty of the grade 1 and 2 tests with student abilities at those grade levels and to increase the separation in difficulty between the K, 1, and 2 tests may be an important future direction for EMSA test development. Better alignment between the overall difficulty of the test and student abilities in grades 1 and 2 may result in less error in the estimation of person

ability, especially for examinees in the upper tails of the distribution of ability at those grade levels. The test is not designed to discriminate among individual students, but this apparent limitation of the test may dampen the ability of the Spring 2016 K–2 EMSA to serve its primary purpose, which was to detect potential achievement differences between students in treatment and control schools in a randomized trial. The comparison of student achievement may be most valid in the grade K subsample but less so with the grades 1 and 2 subsamples.

References

Anderson, D., Kahn, J, & Tindal, G. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Applied Measurement in Education*, 30, 163–177. http://dx.doi.org/10.1080/08957347.2017.1316277

Battauz, M. (2013). IRT test equating in complex linkage plans. *Psychometrika*, 78, 464–480.

Carpenter, T. P., Ansell, E., Franke, M. L., Fennema, E., Weisbeck, L. (1993). Models of problem solving: A study of kindergarten children's problem-solving processes. *Journal for Research in Mathematics Education, 24(5),* 428–441.

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., Franke, M. L., Levi, L., & Empson, S. B. (2015). *Children's mathematics: Cognitively guided instruction* (2nd Ed.). Portsmouth, NH: Heinemann.

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory.* Mason, OH: Cengage Learning.

Dixon, J. K., Larson, M., Leiva, M. A., & Adams, T. L. (2013). *GoMath! Florida.* Orlando, FL: Houghton Mifflin Harcourt.

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)*, 370–378.

Florida Department of Education. (2014). *Mathematics Florida Standards*. Retrieved from http://www.fldoe.org/core/fileparse.php/5390/urlt/0081015-mathfs.pdf

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179–185. doi:10.1007/BF02289447

Kolen, M. J., & Brennan, R.L. (2014). *Test equating, scaling, and linking: methods and practices*, 3rd ed. New York: Springer.

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

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Revelle, W. (2017). *psych: Procedures for personality and psychological research* (Version 1.7.5). Evanston, Illinois: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Resear*ch, 14, 403–414. http://dx.doi.org/10.1207/s15327906mbr1404_2

Schoen, R. C., Anderson, D., Champagne, Z., & Bauduin, C. (2017). Elementary mathematics student assessment: Measuring the performance of grade K, 1, and 2 students in counting, word problems, and computation in fall 2015. (Research Report No. 2017-20.) Tallahassee, FL: Learning Systems Institute, Florida State University. http://dx.doi.org/10.17125/fsu.1522170756

Schoen, R. C., LaVenia, M., Bauduin, C., & Farina, K. (2016a). 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. http://dx.doi.org/10.17125/fsu.1508170543

Schoen, R. C., LaVenia, M., Bauduin, C., & Farina, K. (2016b). Elementary mathematics student assessment: Measuring the performance of grade 1 and 2 students in counting, word problems, and computation in fall 2014. (Research Report No. 2016-04). Tallahassee, FL: Learning Systems Institute, Florida State University. http://dx.doi.org/10.17125/fsu.1508174887

Schoen, R. C., LaVenia, M., Champagne, Z. M., & Farina, K., (2016c). *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: Florida Center for Research in Science, Technology, Engineering, and Mathematics. http://dx.doi.org/10.17125/fsu.1493238156

Schoen, R. C., LaVenia, M., Champagne, Z. M., Farina, K., & Tazaz, A. (2016d). Mathematics performance and cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2015. (Research Report No. 2016–02). Tallahassee, FL: Florida Center for Research in Science, Technology, Engineering, and Mathematics. http://dx.doi.org/10.17125/fsu.1493238666

Turner, E. E., & Celedón-Pattichis, S. (2011). Mathematical problem solving among Latina/o kindergartners: An analysis of opportunities to learn. *Journal of Latinos and Education*, *10(2)*, 146–169.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321–327. http://dx.doi.org/10.1007/BF02293557

Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole numbers concepts and operations. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning.* Reston, VA: National Council of Teachers of Mathematics.

Appendix A. Grade K Test

The form in this appendix is identical to the form used in spring 2016. As a result, no headers or footers are used in this section of the report.

Kindergarten – End of Year
Student Mathematics Assessment

Date: _____

District: _____   School: _____

Teacher: _____

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

What grade are you in?

K	1	2	3	4
●	○	○	○	○

This paper may include some kinds of problems that are new or challenging for you. Don't worry if you can't solve them. You won't be graded on this test, but please try your best!

This page is intentionally left blank.

Fill in the bubble under the shape that is a triangle.

O   O   O   O

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

_____

Yes   No

_____

Yes   No

_____

Yes   No

⬛ = ☐

⬛ = ☐

⬛ = ☐ ⬛

STOP
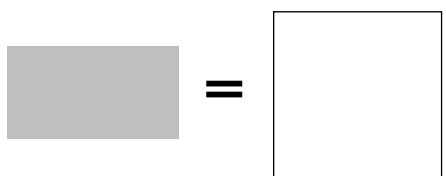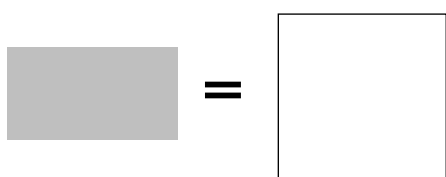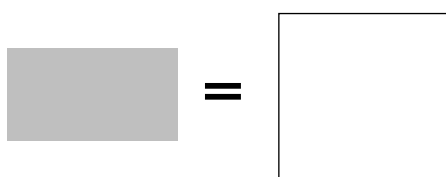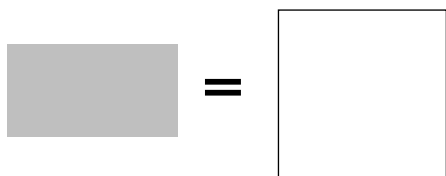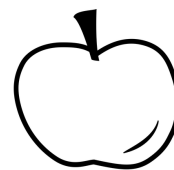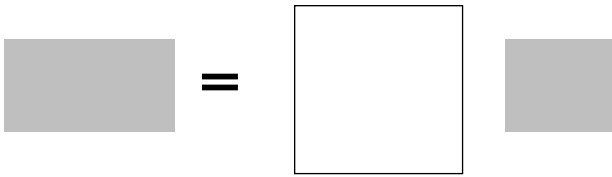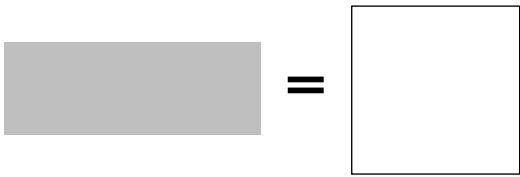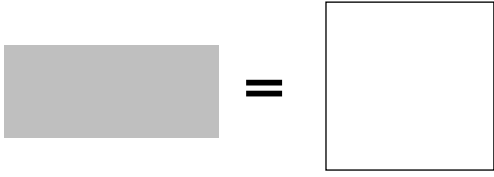
Appendix B. Grade 1 Test

The form in this appendix is identical to the form used in spring 2016. As a result, no headers or footers are used in this section of the report.

First Grade – End of Year
Student Mathematics Assessment

Date: _____

District: _____   School: _____

Teacher: _____

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

What grade are you in?

1	2	3	4	5
●	○	○	○	○

This paper may include some kinds of problems that are new or challenging for you. Don't worry if you can't solve them. You won't be graded on this test, but please try your best!

This page is intentionally left blank.

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

■ □ = ■

■ = □

■ = □

■ = □

▢ = ▢ ▢

▢ = ▢ ▢

STOP
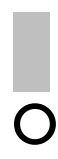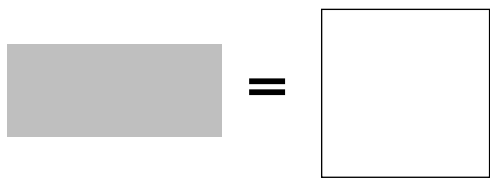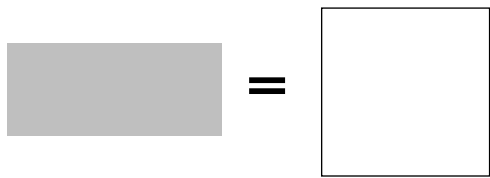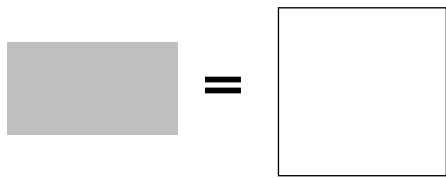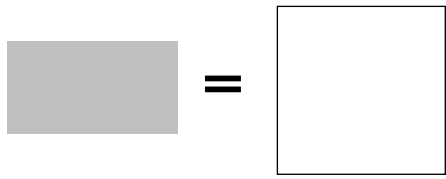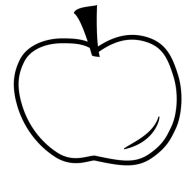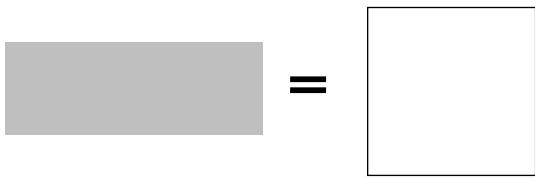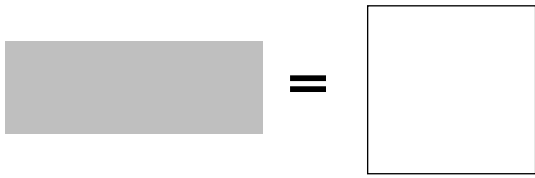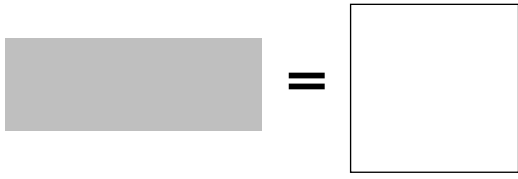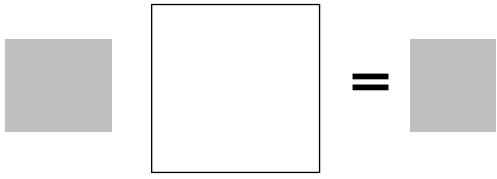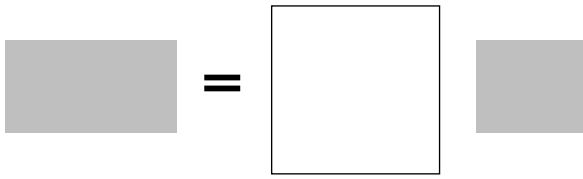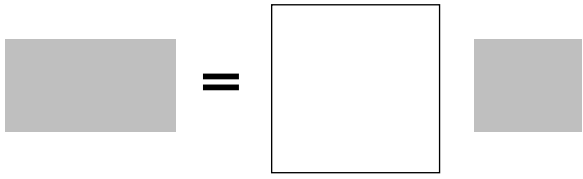
Appendix C. Grade 2 Test

The form in this appendix is identical to the form used in spring 2016. As a result, no headers or footers are used in this section of the report.

Second Grade – End of Year
Student Mathematics Assessment

Date: _____

District: _____   School: _____

Teacher: _____

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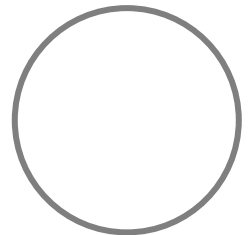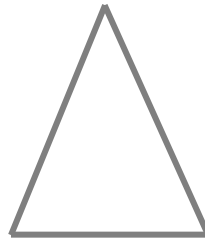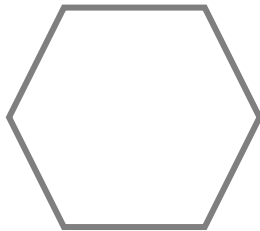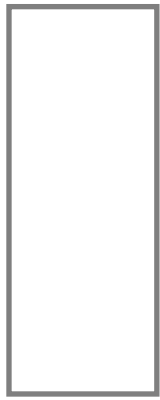
What grade are you in?

1	2	3	4	5
○	●	○	○	○

This paper may include some kinds of problems that are new or challenging for you. Don't worry if you can't solve them. You won't be graded on this test, but please try your best!

This page is intentionally left blank.

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

Yes   No

■ = □

■ = □

■ = □

■ □ = ■

=

=

=

=

⬚⬚⬚ = ☐ ⬚

⬚⬚⬚ = ☐ ⬚

⬚⬚ ☐ = ⬚⬚⬚

STOP

Appendix D. Grade K Administration Guide

The form in this appendix is identical to the form used in spring 2016. As a result, no headers or footers are used in this section of the report.

Foundations for Success in STEM:
Administration Instructions for the Kindergarten End of Year Student Mathematics Assessment

2015–2016

Overview

Thank you for your participation in the *Foundations for Success in STEM* research study. This document will provide you with instructions to follow for the purpose of assessing your mathematics students. The assessment is designed to be administered in a written format with the whole class, but you may administer individually or in small groups as you see fit. Please administer the End of Year Assessment during the allotted testing window listed in the table below. If you cannot administer the assessment during that window, please notify Amanda Tazaz (atazaz@lsi.fsu.edu) and plan to administer it as soon as possible.

You will notice that the End of Year Assessment contains four basic sections: Number Facts, Word Problems, Understanding the Equals Sign, and Computation. Many of the items on the test use a multiple-choice format. We ask that students use pencils to bubble their answers. A script for the teacher to use during administration begins on page 5 of this guide. Please follow the script as closely as possible when you or your surrogate administers the End of Year Assessment. A class roster form is enclosed with this document so that you can provide basic information about the students in your class. Please complete the roster form and include it with the class set of assessments in the envelope provided. The assessments will be picked up as described in the Submitting the End of Year Assessment Materials section on page 4.

End of Year Assessment Window

Student testing will occur according to the following schedule:

School District	Testing Window
District A	April 18th to May 11th, 2016
District B	April 18th to May 11th, 2016
District C	April 18th to May 11th, 2016
District D	April 18th to May 11th, 2016
District E	April 18th to May 11th, 2016
District F	April 18th to May 11th, 2016
District G	April 18th to May 11th, 2016
District H	April 18th to May 11th, 2016
District I	April 18th to May 11th, 2016
District J	April 18th to May 11th, 2016
District K	April 18th to May 11th, 2016
District L	April 18th to May 11th, 2016
District M	April 18th to May 11th, 2016

Materials

The following materials are required for testing:
- Administration Instructions for the Kindergarten End of Year Assessment (this document)
- A test booklet for each student (one per student, provided)
- At least one sharpened pencil for each student

Test Booklets

The students should mark their answers directly in the test booklets. Should you need additional testing materials, please contact Amanda Tazaz (atazaz@lsi.fsu.edu). Remember that these materials are to remain at the school site until the testing window has ended. The materials should be stored in a secure, access-restricted location at all times.

Students to Be Tested
We ask that you administer the End of Year Assessment to students for whom you are the teacher of record. Therefore, even if you teach multiple groups of students, you need only administer it to students who are assigned to your homeroom.

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

Date:
District: School:
Teacher:
Student: Grade:

Before the testing session, *the classroom teacher must enter this information* (date, district name, school name, teacher name, student full name as it appears on official records, and student grade level) on each test booklet for each student to be tested. (Please do not leave this information for students to enter.)

You may administer the End of Year Assessment for the *Foundations for Success in STEM* Study to students on either an individual or a whole-group basis. Please adhere to the following guidelines:
- Ensure all students have testing materials (i.e., test booklet and a sharpened pencil).
- Ensure that students and prelabeled test booklets are properly paired (i.e., that each student receives the test booklet that has his or her name written on it).
- Provide students with a comfortable testing environment.
- Permit students to use mathematics manipulatives during the End of Year Assessment if they would ordinarily be permitted to use manipulatives in your classroom.
- Do not permit any talking or communication between students during testing.
- Please adhere to the End of Year Assessment guidelines and administration instructions.
- Read the test aloud to students as instructed.
- Although the administration script indicates that teachers should read each question two times, read the problem more than two times if necessary. To ensure uniformity of testing, if the problem is read more than two times, read the *entire problem* each time and not simply parts of the problem.

Administering the End of the Year Assessment
We assume that the classroom teacher will administer the End of Year Assessment, but other school personnel (such as paraprofessionals or even substitute teachers) can administer it, provided they follow the assessment protocol as described below.

The testing conditions for the End of Year Assessment should be consistent with the testing conditions for other student assessments administered in the classroom. For example, students should space out their desks or use student "privacy folders" if that is what they would usually do. In addition, if students are normally permitted to use mathematics manipulatives during testing situations, they should be permitted to use them during the End of Year Assessment.

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. To ensure that the students' test

responses are valid, please ensure that appropriate procedures are followed when the End of the Year Assessment is administered. These procedures include:

- Administration of the appropriate test level (Kindergarten assessment for Grade K students, etc.)
- Adherence to the End of the Year Assessment 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.

Testing Time Allocation
Administration of the End of Year Assessment 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 End of the Year Assessment Materials
Upon conclusion of testing, repack the test booklets (both used and unused) in the original packaging. Also, please be sure to include the End of Year Assessment guidelines and administration instruction document and your completed student information sheet in the package. A member of the project will collect the testing materials from your school during the pick-up window listed below. Please have all materials at the front office no later than May 13th, 2016.

The testing materials will be picked up from schools as follows (you will receive an e-mail message before pick-up to ensure the materials are ready in the front office).

School district	Materials due at front office	Pick-up window
District A	May 13th, 2016	May 16th to 20th, 2016
District B	May 13th, 2016	May 16th to 20th, 2016
District C	May 13th, 2016	May 16th to 20th, 2016
District D	May 13th, 2016	May 16th to 20th, 2016
District E	May 13th, 2016	May 16th to 20th, 2016
District F	May 13th, 2016	May 16th to 20th, 2016
District G	May 13th, 2016	May 16th to 20th, 2016
District H	May 13th, 2016	May 16th to 20th, 2016
District I	May 13th, 2016	May 16th to 20th, 2016
District J	May 13th, 2016	May 16th to 20th, 2016
District K	May 13th, 2016	May 16th to 20th, 2016
District L	May 13th, 2016	May 16th to 20th, 2016
District M-	May 13th, 2016	May 16th to 20th, 2016

If you have questions about this process, contact atazaz@lsi.fsu.edu.

Please turn to the next page for the End of Year Assessment Script.

End of Year Assessment Administration Instructions – Kindergarten

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

> You are about to take a math assessment. You will need a pencil.

Verify that all every students has 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. It is your choice if you want to answer the questions or complete the test. Some of these questions may be hard, but don't worry and just try your best.

Ensure that students and pre-labeled test booklets are properly paired (i.e., that each student receives the test booklet that has his or her name written on it). If your students would ordinarily be permitted to use manipulatives and/or scratch paper in this type of situation, ensure that they are available at this time or remind them of your policies for how to access them.

> The first page of the assessment gives the instructions and provides a sample of how you will mark your answers.
>
> The first problems on this assessment are going to ask you to mark your answer choices by filling 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 K, for Kindergarten. Notice how the bubble beneath (below) the K has been filled in for you. You are going to mark your answer choices the same way, by filling in the bubble beneath (below) the answer choice you think is correct.
>
> Turn the page. You should see a pencil in the corner. Let's try this practice one together. It says: Fill in the bubble under the shape that is a triangle. Take your pencil and fill in the bubble below the shape that is a triangle.
>
> The correct answer is the triangle (hold up a test and point to the triangle).

Walk around to ensure that all students have filled in the bubble under the triangle.

For each problem, I would like for you to try hard to figure out which answer is correct. If you are not sure, mark the answer that you think is best.

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, you may use the white space on the paper to work out your answers.

Are there any questions?

Address any questions.

If there are no more questions, turn to the page with the book at the top.

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

The first problem is,

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 dog at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 frog at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 bicycle at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time: ▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓
▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓

Fill in the bubble 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.

For each of the next eight items I am going to read an equation aloud to you. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

The equation at the top of the page is ▓▓▓▓▓▓▓▓▓▓▓▓. Again, ▓▓▓▓▓
▓▓▓▓▓▓▓. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ▓▓▓▓▓▓▓▓▓▓▓. Again, ▓▓▓▓▓▓▓▓▓▓▓▓▓. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ▓▓▓▓▓▓▓▓▓▓▓▓▓▓▓. Again, ▓▓▓▓▓▓▓
▓▓▓▓▓▓▓▓▓. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The last equation on this page is ▓▓▓▓▓▓▓▓▓. Again, ▓▓▓▓▓
▓▓▓▓▓▓. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the balloon at the top.

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

The equation at the top of this page is ▓▓▓▓▓▓▓▓▓▓▓▓. Again, ▓▓▓▓▓▓▓▓▓▓. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ▓▓▓▓▓▓▓▓▓. Again, ▓▓▓▓▓▓▓▓▓▓. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ▨▨▨▨▨. Again, ▨▨▨▨▨. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The last equation on this page is ▨▨▨▨▨▨▨▨. Again, ▨▨
▨▨▨▨▨▨. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

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 at the top.

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

The problem at the top of this page is, ▨▨▨▨▨▨? Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time. ▨▨▨▨▨▨? Fill in the bubble 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.

Look at the next item on this page. It says, ▨▨▨▨▨▨? Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time. ▨▨▨▨▨▨? Fill in the bubble below the answer you think is correct.

When you are finished, put your pencil down.

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

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

Now you are going to work on some problems on your own. The next two pages have some addition and subtraction problems that you will solve at your own pace. Your job is to find out what number goes in the box to make the equation correct. Then you'll write your answer in the box.

Remember to look closely at the symbol to decide if it is an addition or subtraction problem. When I say "begin" you can start answering the questions. When you get to the end of the first page, continue on to the next page until you reach the stop sign at the bottom of that page. Are there any questions?

Address any questions.

BEGIN.

Circulate as students work on the problems. Provide students with ample time to complete the problems. Once you see that students have completed the problems, please end the assessment.

END.

Collect all testing materials.

Appendix E. Grade 1 Administration Guide

The form in this appendix is identical to the form used in spring 2016. As a result, no headers or footers are used in this section of the report.

Foundations for Success in STEM: Administration Instructions for the First Grade End of Year Student Mathematics Assessment

2015–2016

Overview
Thank you for your participation in the *Foundations for Success in STEM* research study. This document will provide you with instructions to follow for the purpose of assessing your mathematics students. The assessment is designed to be administered in a written format with the whole class, but you may administer individually or in small groups as you see fit. Please administer the End of Year Assessment during the allotted testing window listed in the table below. If you cannot administer the assessment during that window, please notify Amanda Tazaz (atazaz@lsi.fsu.edu) and plan to administer it as soon as possible.

You will notice that the End of Year Assessment contains four basic sections: Number Facts, Word Problems, Understanding the Equal Sign, and Computation. Many of the items on the test use a multiple-choice format. We ask that students use pencils to bubble their answers. A script for the teacher to use during administration begins on page 5 of this guide. Please follow the script as closely as possible when you or your surrogate administers the End of Year Assessment. A class roster form is enclosed with this document so that you can provide basic information about the students in your class. Please complete the roster form and include it with the class set of assessments in the envelope provided. The assessments will be picked up as described in the Submitting the End of Year Assessment Materials section on page 4.

End of Year Assessment Window
Student testing will occur according to the following schedule:

School District	Testing Window
District A	April 18th to May 11th, 2016
District B	April 18th to May 11th, 2016
District C	April 18th to May 11th, 2016
District D	April 18th to May 11th, 2016
District E	April 18th to May 11th, 2016
District F	April 18th to May 11th, 2016
District G	April 18th to May 11th, 2016
District H	April 18th to May 11th, 2016
District I	April 18th to May 11th, 2016
District J	April 18th to May 11th, 2016
District K	April 18th to May 11th, 2016
District L	April 18th to May 11th, 2016
District M	April 18th to May 11th, 2016

Materials
The following materials are required for testing:
- Administration Instructions for the Kindergarten End of Year Assessment (this document)
- A test booklet for each student (one per student, provided)
- At least one sharpened pencil for each student

Test Booklets
The students should mark their answers directly in the test booklets. Should you need additional testing materials, please contact Amanda Tazaz (atazaz@lsi.fsu.edu). Remember that these materials are to remain at the school site until the testing window has ended. The materials should be stored in a secure, access-restricted location at all times.

Students to Be Tested
We ask that you administer the End of Year Assessment to students for whom you are the teacher of record. Therefore, even if you teach multiple groups of students, you need only administer it to students who are assigned to your homeroom.

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

Date:
District: School:
Teacher:
Student: Grade:

Before the testing session, *the classroom teacher must enter this information* (date, district name, school name, teacher name, student full name as it appears on official records, and student grade level) on each test booklet for each student to be tested. (Please do not leave this information for students to enter.)

You may administer the End of Year Assessment for the *Foundations for Success in STEM* Study to students on either an individual or a whole-group basis. Please adhere to the following guidelines:
- Ensure all students have testing materials (i.e., test booklet and a sharpened pencil).
- Ensure that students and prelabeled test booklets are properly paired (i.e., that each student receives the test booklet that has his or her name written on it).
- Provide students with a comfortable testing environment.
- Permit students to use mathematics manipulatives during the End of Year Assessment if they would ordinarily be permitted to use manipulatives in your classroom.
- Do not permit any talking or communication between students during testing.
- Please adhere to the End of Year Assessment guidelines and administration instructions.
- Read the test aloud to students as instructed.
- Although the administration script indicates that teachers should read each question two times, read the problem more than two times if necessary. To ensure uniformity of testing, if the problem is read more than two times, read the *entire problem* and not simply parts of the problem.

Administering the End of the Year Assessment
We assume that the classroom teacher will administer the End of Year Assessment, but other school personnel (such as paraprofessionals or even substitute teachers) can administer it, provided they follow the assessment protocol as described below.

The testing conditions for the End of Year Assessment should be consistent with the testing conditions for other student assessments administered in the classroom. For example, students should space out their desks or use student "privacy folders" if that is what they would usually do. In addition, if students are normally permitted to use mathematics manipulatives during testing situations, they should be permitted to use them during the End of Year Assessment.

Avoid reading problems or answering student questions in a way that may offer clues to the correct answers. Student responses should reflect their current math knowledge. To ensure that the students' test responses are valid, please ensure that appropriate procedures are followed when the End of the Year Assessment is administered. These procedures include:

- Administration of the appropriate test level (Kindergarten assessment for Grade K students, etc.)
- Adherence to the End of the Year Assessment 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.

Testing Time Allocation
Administration of the End of Year Assessment 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 End of the Year Assessment Materials
Upon conclusion of testing, repack the test booklets (both used and unused) in the original packaging. Also, please be sure to include the End of Year Assessment guidelines and administration instruction document and your completed student information sheet in the package. A member of the project will collect the testing materials from your school during the pick-up window listed below. Please have all materials at the front office no later than May 13th, 2016.

The testing materials will be picked up from schools as follows (you will receive an e-mail message before pick-up to ensure the materials are ready in the front office).

School district	Materials due at front office	Pick-up window
District A	May 13th, 2016	May 16th to 20th, 2016
District B	May 13th, 2016	May 16th to 20th, 2016
District C	May 13th, 2016	May 16th to 20th, 2016
District D	May 13th, 2016	May 16th to 20th, 2016
District E	May 13th, 2016	May 16th to 20th, 2016
District F	May 13th, 2016	May 16th to 20th, 2016
District G	May 13th, 2016	May 16th to 20th, 2016
District H	May 13th, 2016	May 16th to 20th, 2016
District I	May 13th, 2016	May 16th to 20th, 2016
District J	May 13th, 2016	May 16th to 20th, 2016
District K	May 13th, 2016	May 16th to 20th, 2016
District L	May 13th, 2016	May 16th to 20th, 2016
District M	May 13th, 2016	May 16th to 20th, 2016

If you have questions about this process, contact atazaz@lsi.fsu.edu.

Please turn to the next page for the End of Year Assessment Script.

End of Year Assessment Administration Instructions – First Grade

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

You are about to take a math assessment. You will need a pencil.

Verify that every student has 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. It is your choice if you want to answer the questions or complete the test. Some of these questions may be hard, but don't worry and just try your best.

Ensure that students and prelabeled test booklets are properly paired (i.e., that each student receives the test booklet that has his or her name written on it). If your students would ordinarily be permitted to use manipulatives and/or scratch paper in this type of situation, ensure that they are available at this time or remind them of your policies for how to access them.

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

The first problems on this assessment are going to ask you to mark your answer choices by filling in the bubble 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 one, for first grade. Notice how the bubble below the one has been filled in for you. You are going to mark your answer choices the same way, by filling in the bubble below the answer choice you think is correct.

For each problem, I would like for you to try hard to figure out which answer is correct. If you are not sure, mark the answer that you think is best.

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, you may use the white space on the paper to work out your answers.

Are there any questions?

Address any questions.

If there are no more questions, turn to the page with the dog at the top.

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

The first problem is,

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 frog at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 bicycle at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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.

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 pencil at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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.

For each of the next eight items I am going to read an equation aloud to you. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

The equation at the top of the page is ░░░░░░░░░░░░░░░░. Again, ░░░░░░ ░░░░░░░. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ░░░░░░░░░░░░░░. Again, ░░░░░░░░░░░░░░░. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ░░░░░░░░░░░░░░░░░░. Again, ░░░░░░░ ░░░░░░░░░. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The last equation on this page is ░░░░░░░░░░░░░░. Again, ░░░░░░ ░░░░░░░. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the balloon at the top.

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

The equation at the top of this page is [   ]. Again, [   ]. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is [   ]. Again, [   ]. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is [   ]. Again, [   ]. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The last equation on this page is [   ]. Again, [   ]. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

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 at the top.

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

The problem at the top of this page is, [redacted]? Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time. [redacted]? Fill in the bubble 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.

Look at the next item on this page. It says, [redacted]? Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time. [redacted]? Fill in the bubble 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.

Look at the last item on this page. It says, [redacted]? Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time. [redacted]? Fill in the bubble 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 apple at the top.

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

Now you are going to work on some problems on your own. The next three pages have some addition and subtraction problems that you will solve at your own pace. Your job is to find out what number goes in the box to make the equation correct. Then you'll write your answer in the box.

Remember to look closely at the symbol to decide if it is an addition or subtraction problem. When I say "begin" you can start answering the questions. When you get to the end of the first page, continue on to the next few pages until you reach the stop sign at the bottom of the last page. Are there any questions?

Address any questions.

BEGIN.

Circulate as students work on the problems. Provide students with ample time to complete the problems. Once you see that students have completed the problems, please end the assessment.

END.

Collect all testing materials.

Appendix F. Grade 2 Administration Guide

The form in this appendix is identical to the form used in spring 2016. As a result, no headers or footers are used in this section of the report.

Foundations for Success in STEM:
Administration Instructions for the Second Grade
End of Year Student Mathematics Assessment

2015–2016

Overview

Thank you for your participation in the *Foundations for Success in STEM* research study. This document will provide you with instructions to follow for the purpose of assessing your mathematics students. The assessment is designed to be administered in a written format with the whole class, but you may administer individually or in small groups as you see fit. Please administer the End of Year Assessment during the allotted testing window listed in the table below. If you cannot administer the assessment during that window, please notify Amanda Tazaz (atazaz@lsi.fsu.edu) and plan to administer it as soon as possible.

You will notice that the End of Year Assessment contains four basic sections: Number Facts, Word Problems, Understanding the Equal Sign, and Computation. Many of the items on the test use a multiple-choice format. We ask that students use pencils to bubble their answers. A script for the teacher to use during administration begins on page 5 of this guide. Please follow the script as closely as possible when you or your surrogate administers the End of Year Assessment. A class roster form is enclosed with this document so that you can provide basic information about the students in your class. Please complete the roster form and include it with the class set of assessments in the envelope provided. The assessments will be picked up as described in the Submitting the End of Year Assessment Materials section on page 4.

End of Year Assessment Window

Student testing will occur according to the following schedule:

School District	Testing Window
District A	April 18th to May 11th, 2016
District B	April 18th to May 11th, 2016
District C	April 18th to May 11th, 2016
District D	April 18th to May 11th, 2016
District E	April 18th to May 11th, 2016
District F	April 18th to May 11th, 2016
District G	April 18th to May 11th, 2016
District H	April 18th to May 11th, 2016
District I	April 18th to May 11th, 2016
District J	April 18th to May 11th, 2016
District K	April 18th to May 11th, 2016
District L	April 18th to May 11th, 2016
District M	April 18th to May 11th, 2016

Materials

The following materials are required for testing:
- Administration Instructions for the Kindergarten End of Year Assessment (this document)
- A test booklet for each student (one per student, provided)
- At least one sharpened pencil for each student

Test Booklets

The students should mark their answers directly in the test booklets. Should you need additional testing materials, please contact Amanda Tazaz (atazaz@lsi.fsu.edu). Remember that these materials are to remain at the school site until the testing window has ended. The materials should be stored in a secure, access-restricted location at all times.

Students to Be Tested
We ask that you administer the End of Year Assessment to students for whom you are the teacher of record. Therefore, even if you teach multiple groups of students, you need only administer it to students who are assigned to your homeroom.

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

Date:
District: School:
Teacher:
Student: Grade:

Before the testing session, *the classroom teacher must enter this information* (date, district name, school name, teacher name, student full name as it appears on official records, and student grade level) on each test booklet for each student to be tested. (Please do not leave this information for students to enter.)

You may administer the End of Year Assessment for the *Foundations for Success in STEM* Study to students on either an individual or a whole-group basis. Please adhere to the following guidelines:
- Ensure all students have testing materials (i.e., test booklet and a sharpened pencil).
- Ensure that students and prelabeled test booklets are properly paired (i.e., that each student receives the test booklet that has his or her name written on it).
- Provide students with a comfortable testing environment.
- Permit students to use mathematics manipulatives during the End of Year Assessment if they would ordinarily be permitted to use manipulatives in your classroom.
- Do not permit any talking or communication between students during testing.
- Please adhere to the End of Year Assessment guidelines and administration instructions.
- Read the test aloud to students as instructed.
- Although the administration script indicates that teachers should read each question two times, read the problem more than two times if necessary. To ensure uniformity of testing, if the problem is read more than two times, read the *entire problem* and not simply parts of the problem.

Administering the End of the Year Assessment
We assume that the classroom teacher will administer the End of Year Assessment, but other school personnel (such as paraprofessionals or even substitute teachers) can administer it, provided they follow the assessment protocol as described below.

The testing conditions for the End of Year Assessment 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. In addition, if students are normally permitted to use mathematics manipulatives during testing situations, they should be permitted to use them during the End of Year Assessment.

Avoid reading problems or answering student questions in a way that may offer clues to the correct answers. Student responses should reflect their current math knowledge. To ensure that the students' test responses are valid, please ensure that appropriate procedures are followed when the End of the Year Assessment is administered. These procedures include:

- Administration of the appropriate test level (Kindergarten assessment for Grade K students, etc.)
- Adherence to the End of the Year Assessment 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.

Testing Time Allocation
Administration of the End of Year Assessment 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 End of the Year Assessment Materials
Upon conclusion of testing, repack the test booklets (both used and unused) in the original packaging. Also, please be sure to include the End of Year Assessment guidelines and administration instruction document and your completed student information sheet in the package. A member of the project will collect the testing materials from your school during the pick-up window listed below. Please have all materials at the front office no later than May 13th, 2016.

The testing materials will be picked up from schools as follows (you will receive an e-mail message before pick-up to ensure the materials are ready in the front office).

School District	Materials due to front office	Pick-up Window
District A	May 13th, 2016	May 16th to 20th, 2016
District B	May 13th, 2016	May 16th to 20th, 2016
District C	May 13th, 2016	May 16th to 20th, 2016
District D	May 13th, 2016	May 16th to 20th, 2016
District E	May 13th, 2016	May 16th to 20th, 2016
District F	May 13th, 2016	May 16th to 20th, 2016
District G	May 13th, 2016	May 16th to 20th, 2016
District H	May 13th, 2016	May 16th to 20th, 2016
District I	May 13th, 2016	May 16th to 20th, 2016
District J	May 13th, 2016	May 16th to 20th, 2016
District K	May 13th, 2016	May 16th to 20th, 2016
District L	May 13th, 2016	May 16th to 20th, 2016
District M	May 13th, 2016	May 16th to 20th, 2016

If you have questions about this process, contact atazaz@lsi.fsu.edu.

Please turn to the next page for the End of Year Assessment Script.

End of Year Assessment Administration Instructions – Second Grade

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

You are about to take a math assessment. You will need a pencil.

Verify that every student has 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. It is your choice if you want to answer the questions or complete the test. Some of these questions may be hard, but don't worry and just try your best.

Ensure that students and prelabeled test booklets are properly paired (i.e., that each student receives the test booklet that has his or her name written on it). If your students would ordinarily be permitted to use manipulatives and/or scratch paper in this type of situation, ensure that they are available at this time or remind them of your policies for how to access them.

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

The first problems on this assessment are going to ask you to mark your answer choices by filling in the bubble 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 two, for second grade. Notice how the bubble below the two has been filled in for you. You are going to mark your answer choices the same way, by filling in the bubble below the answer choice you think is correct.

For each problem, I would like for you to try hard to figure out which answer is correct. If you are not sure, mark the answer that you think is best.

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, you may use the white space on the paper to work out your answers.

Are there any questions?

Address any questions.

If there are no more questions, turn to the page with the bicycle at the top.

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

The first problem is,

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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.

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 pencil at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 dog at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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 frog at the top.

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

Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time:

Fill in the bubble 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.

For each of the next eight items I am going to read an equation aloud to you. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

The equation at the top of the page is ⬚⬚⬚⬚. Again, ⬚⬚⬚⬚. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ⬚⬚⬚⬚. Again, ⬚⬚⬚⬚. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ⬚⬚⬚⬚. Again, ⬚⬚⬚⬚. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The last equation on this page is ⬚⬚⬚⬚. Again, ⬚⬚⬚⬚. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the balloon at the top.

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

The equation at the top of this page is ░░░░░░░░░░░░░░░░░░. Again, ░░░░░░░░░░░░░░░. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ░░░░░░░░░░░░. Again, ░░░░░░░░░░. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The next equation is ░░░░░░░░. Again, ░░░░░░░. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

The last equation on this page is ░░░░░░░░░░░░░░░░. Again, ░░░░░░░░░░░░░░░░. If the equation is correct, circle the word yes. If the equation is not correct, circle the word no.

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 at the top.

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

The problem on this page is ⬚⬚⬚⬚⬚⬚⬚⬚⬚⬚⬚? Fill in the bubble below the answer you think is correct.

I am going to read the problem one more time. ⬚⬚⬚⬚⬚⬚⬚⬚⬚⬚? Fill in the bubble 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 apple at the top.

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

Now you are going to work on some problems on your own. The next three pages have some addition and subtraction problems that you will solve at your own pace. Your job is to find out what number goes in the box to make the equation correct. Then you'll write your answer in the box.

Remember to look closely at the symbol to decide if it is an addition or subtraction problem. When I say "begin" you can start answering the questions. When you get to the end of the first page, continue on to the next few pages until you reach the stop sign at the bottom of the last page. Are there any questions?

Address any questions.

BEGIN.

Circulate as students work on the problems. Provide students with ample time to complete the problems. Once you see that students have completed the problems, please end the assessment.

END.

Collect all testing materials.

Appendix G. Scoring Key

Table G.1. Scoring Key for All Items

Item	Item description	Data entry	Correct response(s)
GKi16		Record student's response, DNS, MR, or UI	4
GKG1i17		Record student's response, DNS, MR, or UI	11
GKG1i18_G2i15		Record student's response, DNS, MR, or UI	14
GKi19		Record student's response, DNS, MR, or UI	3
GKi20_G1i19_G2i		Record student's response, DNS, MR, or UI	05, 5
G1i20_G2i17		Record student's response, DNS, MR, or UI	07, 7
G1i21_G2i18		Record student's response, DNS, MR, or UI	08, 8
GKG1G2i8		Yes, No, DNS, MR, or UI	Yes
GKG1G2i10		Yes, No, DNS, MR, or UI	No
GKG1G2i13		Yes, No, DNS, MR, or UI	Yes
GKi22_G1i25_G2i		Record student's response, DNS, MR, or UI	5
G1i26_G2i24		Record student's response, DNS, MR, or UI	1
GKi2		Record student's response, DNS, MR or UI	7
GKi4_G1i2		Record student's response, DNS, MR, or UI	15
GKi3_G1i1		Record student's response, DNS, MR, or UI	12
GKi5_G1i3_G2i1		Record student's response, DNS, MR, or UI	5
G1i4_G2i2		Record student's response, DNS, MR, or UI	9
G1i5_G2i3		Record student's response, DNS, MR, or UI	6
G2i4		Record student's response, DNS, MR, or UI	25
G2i5		Record student's response, DNS, MR, or UI	140
GKG1G2i6		Yes, No, DNS, MR, or UI	Yes
GKG1G2i7		Yes, No, DNS, MR, or UI	No
GKG1G2i12		Yes, No, DNS, MR, or UI	Yes
GKG1G2i9		Yes, No, DNS, MR, or UI	Yes
GKG1G2i11		Yes, No, DNS, MR, or UI	Yes
GKG1i14		Record student's response, DNS, MR, or UI	29
GKG1i15		Record student's response, DNS, MR, or UI	26
G1i16_G2i14		Record student's response, DNS, MR, or UI	37
GKi21_G1i22		Record student's response, DNS, MR, or UI	35
G1i23_G2i19		Record student's response, DNS, MR, or UI	99
G1i24_G2i20		Record student's response, DNS, MR, or UI	02, 2
G2i21		Record student's response, DNS, MR, or UI	006, 06, 6, 6+
G2i22		Record student's response, DNS, MR, or UI	686
G2i25		Record student's response, DNS, MR, or UI	101

Appendix H. Results of Initial Screening

Appendix H contains results of various analyses performed during the item screening process.

H.1 Item-level Statistics

Tables H.1, H.2, and H.3 present point estimates for the various classical test theory (CTT)- and item-response theory (IRT)-based statistics. Items with statistics missing in the IRT-based statistics columns were removed during the initial screening or during review of the IRT-based model data.

Table H.1. Item Statistics for the Grade K Test Based on the Grade K Sample (n = 950)

| Item | Item description | CTT-based statistics | | IRT-based statistics | | | |
| | | | | Vertical scale | | Within-grade-level scale | |
Item	Item description	PC (se)	PB	Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
GKi2		.86 (.011)	.39	3.01 (.173)	−1.84 (.171)	1.53 (.173)	−2.52 (.171)
GKi3_G1i1		.82 (.013)	.44	3.20 (.170)	−1.67 (.153)	1.63 (.170)	−2.13 (.153)
GKi4_G1i2		.45 (.016)	.56	2.61 (.125)	−.90 (.088)	1.33 (.125)	.27 (.088)
GKi5_G1i3_G2i1		.23 (.014)	.46	2.25 (.123)	−.34 (.106)	1.15 (.123)	1.48 (.106)
GKG1G2i6		*.92 (.009)*	*.31*	–	–	–	–
GKG1G2i7		*.90 (.010)*	*.31*	–	–	–	–
GKG1G2i8		.44 (.016)	.41	.94 (.038)	−.75 (.069)	.48 (.038)	.24 (.069)
GKG1G2i9		.49 (.016)	.38	.94 (.038)	−.94 (.068)	.48 (.038)	.06 (.068)
GKG1G2i10		.71 (.015)	.28	.94 (.038)	−1.98 (.075)	.48 (.038)	−.92 (.075)
GKG1G2i11		.50 (.016)	.36	.94 (.038)	−1.00 (.068)	.48 (.038)	<.01 (.068)
GKG1G2i12		.61 (.016)	.41	.94 (.038)	−1.50 (.070)	.48 (.038)	−.47 (.070)
GKG1G2i13		.39 (.016)	.32	.94 (.038)	−.49 (.070)	.48 (.038)	.48 (.070)
GKG1i14		.44 (.016)	.56	2.82 (.132)	−.90 (.091)	1.44 (.132)	.30 (.091)
GKG1i15		.44 (.016)	.43	1.73 (.097)	−.83 (.077)	.88 (.097)	.30 (.077)
GKi16		.90 (.010)	.40	3.99 (.241)	−1.85 (.265)	2.03 (.241)	−3.39 (.265)
GKG1i17		.69 (.015)	.57	4.91 (.250)	−1.30 (.159)	2.50 (.250)	−1.48 (.159)
GKG1i18_G2i15		.57 (.016)	.54	3.62 (.170)	−1.13 (.105)	1.84 (.170)	−.46 (.105)
GKi19		.58 (.016)	.53	2.71 (.129)	−1.17 (.091)	1.38 (.129)	−.44 (.091)
GKi20_G1i19_G2i16		.45 (.016)	.54	2.81 (.133)	−.91 (.090)	1.43 (.133)	.25 (.090)
GKi21_G1i22		.38 (.016)	.56	3.59 (.170)	−.80 (.108)	1.83 (.170)	.74 (.108)
GKi22_G1i25_G2i23		*.14 (.011)*	*.39*	–	–	–	–

Note. CTT = classical test theory; IRT = item response theory; PC = proportion correct; PB = point biserial; Diff = Difficulty; Discrim = discrimination. Italicized items were removed as a result of initial screening.

Table H.2. Item Statistics for the Grade 1 Test Based on the Grade 1 Sample (n = 1,821)

| | | CTT-based statistics | | IRT-based statistics | | | |
| | | | | Vertical scale | | Within-grade-level scale | |
Item	Item description	PC (se)	PB	Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
GKi3_G1i1		.96 (.005)	.25	1.25 (.170)	−2.99 (.195)	1.25 (.170)	−3.74 (.195)
GKi4_G1i2		.66 (.011)	.62	1.90 (.115)	−.58 (.090)	1.90 (.115)	−1.10 (.090)
GKi5_G1i3_G2i1		.62 (.011)	.61	1.76 (.105)	−.46 (.079)	1.76 (.105)	−.81 (.079)
G1i4_G2i2		.62 (.011)	.60	1.62 (.098)	−.48 (.075)	1.62 (.098)	−.79 (.075)
G1i5_G2i3		*.48 (.012)*	*.54*	–	–	–	–
GKG1G2i6		*.99 (.003)*	*.17*	–	–	–	–
GKG1G2i7		*.99 (.003)*	*.19*	–	–	–	–
GKG1G2i8		.64 (.011)	.61	1.73 (.051)	−.54 (.074)	1.73 (.051)	−.94 (.074)
GKG1G2i9		.78 (.010)	.49	1.73 (.051)	−1.12 (.082)	1.73 (.051)	−1.93 (.082)
GKG1G2i10		.83 (.009)	.42	1.73 (.051)	−1.34 (.088)	1.73 (.051)	−2.32 (.088)
GKG1G2i11		.73 (.010)	.49	1.73 (.051)	−.91 (.068)	1.73 (.051)	−1.57 (.078)
GKG1G2i12		.81 (.009)	.48	1.73 (.051)	−1.26 (.085)	1.73 (.051)	−2.18 (.085)
GKG1G2i13		.57 (.012)	.56	1.73 (.051)	−.28 (.072)	1.73 (.051)	−.48 (.072)
GKG1i14		.75 (.010)	.58	1.93 (.126)	−.94 (.108)	1.93 (.126)	−1.82 (.108)
GKG1i15		.76 (.010)	.52	1.58 (.107)	−1.08 (.094)	1.58 (.107)	−1.70 (.094)
G1i16_G2i14		.71 (.011)	.58	1.73 (.109)	−.80 (.089)	1.73 (.109)	−1.37 (.089)
GKG1i17		.93 (.006)	.34	1.49 (.154)	−2.26 (.172)	1.49 (.154)	−3.37 (.172)
GKG1i18		.85 (.008)	.34	.93 (.092)	−2.19 (.088)	.93 (.092)	−2.04(.088)
GKi20_G1i19_G2i16		.80 (.009)	.46	1.29 (.098)	−1.42 (.090)	1.29 (.098)	−1.82 (.090)
G1i20_G2i17		.74 (.010)	.44	1.01 (.079)	−1.23 (.069)	1.01 (.079)	−1.24 (.069)
G1i21_G2i18		.77 (.010)	.48	1.32 (.095)	−1.22 (.085)	1.32 (.095)	−1.61 (.085)
GKi21_G1i22		.81 (.009)	.51	1.56 (.112)	−1.29 (.104)	1.56 (.112)	−2.02 (.104)
G1i23_G2i19		.54 (.012)	.51	1.12 (.074)	−.20 (.059)	1.12 (.074)	−.22 (.059)
G1i24_G2i20		.28 (.011)	.39	.86 (.069)	1.23 (.062)	.86 (.069)	1.06 (.062)
GKi22_G1i25_G2i23		.40 (.011)	.62	1.73 (.051)	.33 (.073)	1.73 (.051)	.58 (.073)
G1i26_G2i24		.38 (.011)	.65	1.73 (.051)	.42 (.073)	1.73 (.051)	.72 (.073)

Note. CTT = classical test theory; IRT = item response theory; PC = proportion correct; PB = point biserial; Diff = Difficulty; Discrim = discrimination. Italicized items were removed as a result of initial screening.

Table H.3. Item Statistics for the Grade 2 Test Based on the Grade 2 Sample (n = 1,764)

| Item | Item description | CTT–based statistics | | IRT-based statistics | | | |
| | | | | Vertical scale | | Within-grade-level scale | |
		PC (se)	PB	Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
GKi5_G1i3_G2i1		.82 (.009)	.50	1.45 (.114)	−.94 (.107)	1.53 (.114)	−2.11 (.107)
G1i4_G2i2		.74 (.010)	.53	1.33 (.096)	−.57 (.083)	1.39 (.096)	−1.43 (.083)
G1i5_G2i3		.60 (.012)	.57	1.38 (.089)	.06 (.070)	1.45 (.089)	−.61 (.070)
G2i4		.79 (.010)	.55	1.61 (.118)	−.70 (.107)	1.70 (.118)	−1.95 (.107)
G2i5		.56 (.012)	.58	1.34 (.086)	.24 (.067)	1.41 (.086)	−.36 (.067)
GKG1G2i6		*.99 (.002)*	*.14*	–	–	–	–
GKG1G2i7		*.99 (.002)*	*.15*	–	–	–	–
GKG1G2i8		.66 (.011)	.64	2.18 (.066)	−.08 (.090)	2.29 (.066)	−1.28 (.090)
GKG1G2i9		.85 (.008)	.48	2.18 (.066)	−.89 (.107)	2.29 (.066)	−3.05 (.107)
GKG1G2i10		.79 (.010)	.53	2.18 (.066)	−.59 (.099)	2.29 (.066)	−2.39 (.099)
GKG1G2i11		.78 (.010)	.49	2.18 (.066)	−.54 (.098)	2.29 (.066)	−2.29 (.098)
GKG1G2i12		.86 (.008)	.44	2.18 (.066)	−.93 (.108)	2.29 (.066)	−3.12 (.108)
GKG1G2i13		.62 (.012)	.65	2.18 (.066)	.04 (.088)	2.29 (.066)	−1.00 (.088)
G1i16_G2i14		.84 (.009)	.52	1.75 (.137)	−.92 (.133)	1.84 (.137)	−2.50 (.133)
GKG1i18_G2i15		.92 (.006)	.27	.85 (.117)	−2.75 (.119)	.89 (.117)	−2.77 (.119)
GKi20_G1i19		.91 (.007)	.36	1.17 (.127)	−1.92 (.133)	1.23 (.127)	−2.85 (.133)
G1i20_G2i17		.85 (.009)	.38	.95 (.095)	−1.64 (.091)	1.00 (.095)	−2.04 (.091)
G1i21_G2i18		.90 (.007)	.44	1.59 (.147)	−1.40 (.157)	1.68 (.147)	−3.04(.157)
G1i23_G2i19		.72 (.011)	.50	1.18 (.087)	−.52 (.075)	1.24 (.087)	−1.21 (.075)
G1i24_G2i20		.56 (.012)	.50	.99 (.072)	.17 (.059)	1.04 (.072)	−.33 (.059)
G2i21		.33 (.011)	.54	1.37 (.089)	1.21 (.073)	1.44 (.089)	.96 (.073)
G2i22		.54 (.012)	.45	.84 (.066)	.28 (.056)	.88 (.066)	−.19 (.056)
GKi22_G1i25_G2i23		.51 (.012)	.72	2.18 (.066)	.45 (.086)	2.29 (.066)	−.13 (.086)
G1i26_G2i24		.49 (.012)	.70	2.18 (.066)	.50 (.086)	2.29 (.066)	−.02 (.086)
G2i25		.37 (.011)	.65	2.18 (.066)	.93 (.088)	2.29 (.066)	.93 (.088)

Note. CTT = classical test theory; IRT = item response theory; PC = proportion correct; PB = point biserial; Diff = Difficulty; Discrim = discrimination. Italicized items were removed as a result of initial screening.

H.2 Spaghetti Plots

Figures H.1, H.2, and H.3 contain spaghetti plots based on all of the items on the tests using a CTT-based approach with some smoothing. The shapes of most of the trace lines appear satisfactory, but several items corresponded to trace lines with u-shaped curves. Those items were further scrutinized during the initial screening.
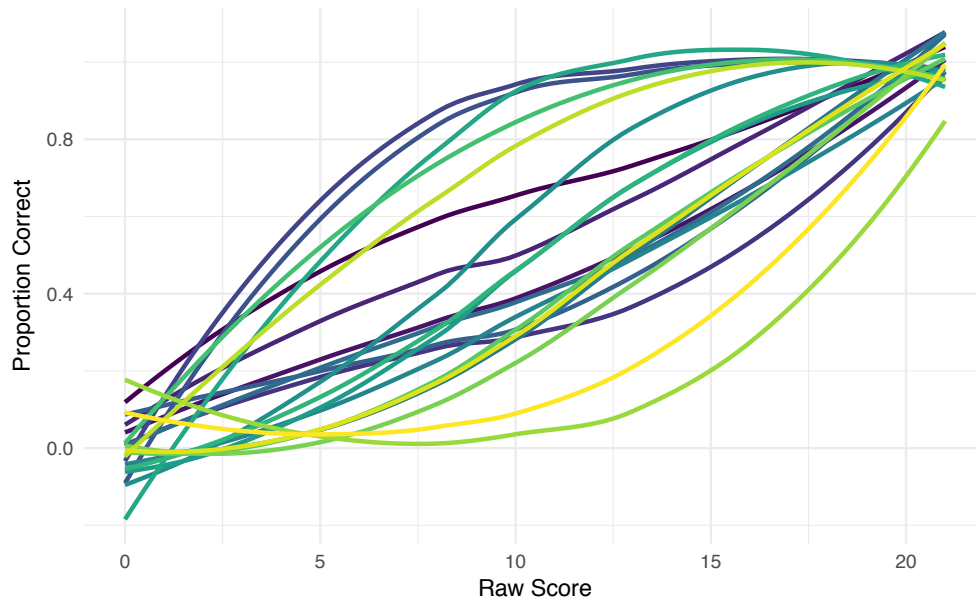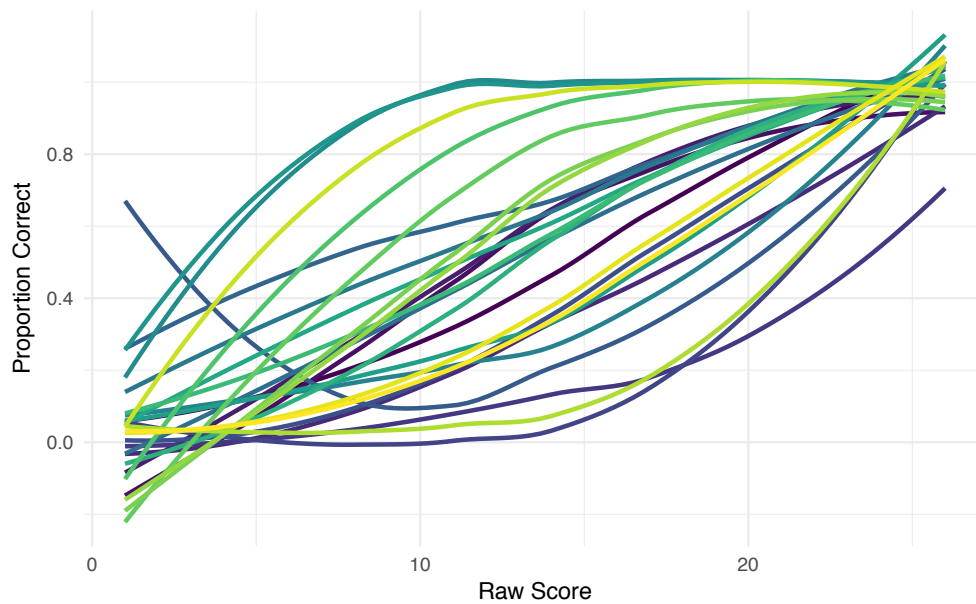

Figure H.1. Grade K spaghetti plot.
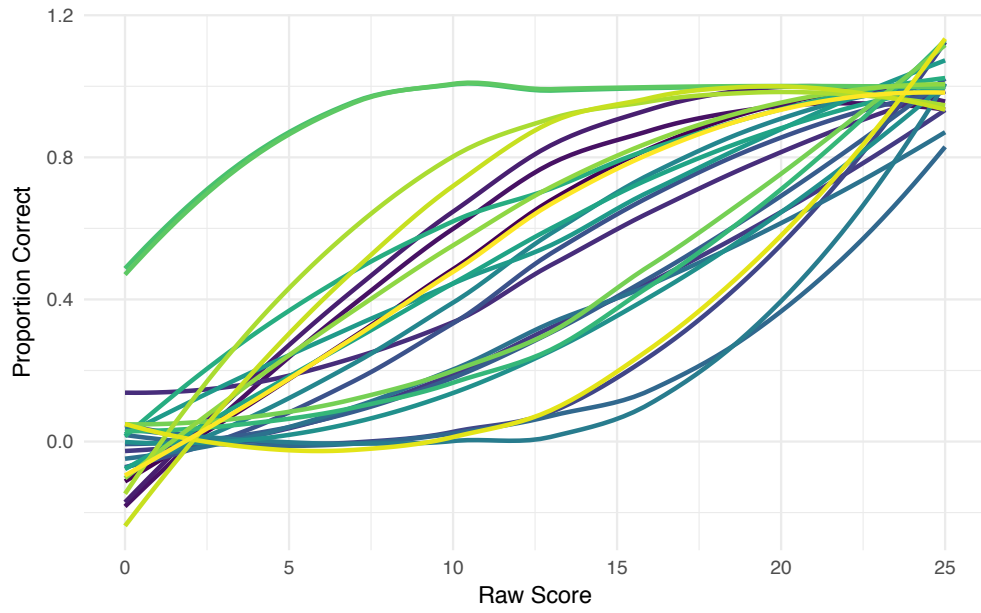

Figure H.2. Grade 1 spaghetti plot.

Figure H.3. Grade 2 spaghetti plot.

Appendix I. Most Common Incorrect Responses for Each Item

Table I.1. Proportion of Grade K Student Responses by Item

Item	Item description	Correct response Response (%)	Most frequent incorrect responses Response (%)	Response (%)	Response (%)	Response (%)
GKi2		7 (.86)	1 (.04)	6 (.03)	4 (.03)	3 (.02)
GKi3_G1i1		12 (.82)	10 (.08)	7 (.04)	5 (.02)	2 (.02)
GKi4_G1i2		15 (.45)	8 (.37)	5 (.06)	20 (.05)	3 (.04)
GKi5_G1i3_G2i1		5 (.23)	13 (.35)	9 (.26)	4 (.07)	6 (.06)
GKG1G2i6		Y (.92)	N (.06)	UI (.01)	DNS (.01)	MR (<.01)
GKG1G2i7		N (.90)	Y (.08)	DNS (.01)	UI (.01)	MR (<.01)
GKG1G2i8		Y (.44)	N (.53)	UI (.02)	DNS (.01)	—
GKG1G2i9		Y (.49)	N (.48)	DNS (.02)	UI (.01)	MR (<.01)
GKG1G2i10		N (.71)	Y (.28)	DNS (.01)	UI (.01)	—
GKG1G2i11		Y (.50)	N (.48)	DNS (.01)	UI (.01)	MR (<.01)
GKG1G2i12		Y (.61)	N (.36)	DNS (.02)	UI (.01)	—
GKG1G2i13		Y (.39)	N (.58)	DNS (.02)	UI (.01)	MR (<.01)
GKG1i14		29 (.44)	2 (.19)	40 (.13)	31 (.12)	20 (.10)
GKG1i15		26 (.44)	20 (.17)	17 (.17)	25 (.13)	6 (.06)
GKi16		4 (.90)	2 (.02)	3 (.02)	DNS (.01)	5 (.01)
GKG1i17		11 (.69)	10 (.09)	6 (.05)	7 (.03)	12 (.03)
GKG1i18_G2i15		14 (.57)	13 (.05)	16 (.04)	15 (.04)	9 (.03)
GKi19		3 (.58)	9 (.17)	4 (.04)	8 (.04)	6 (.04)
GKi20_G1i19		5 (.45)	19 (.10)	6 (.05)	3 (.04)	4 (.04)
GKi21_G1i22		35 (.38)	30 (.05)	34 (.04)	15 (.04)	10 (.03)
GKi22_G1i25_G2i23		5 (.14)	9 (.39)	13 (.08)	4 (.05)	8 (.05)

Note. n = 950 valid grade K tests conducted. Italicized items were removed as a result of initial screening. Those that were not answered were recorded as "DNS." Item responses that were unclear were recorded as "UI." GKi1 was not scored and was used as a practice question.

Table I.2. Proportion of Grade 1 Student Responses by Item

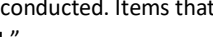| Item	Item description	Correct response Response (%)	Most frequent incorrect responses Response (%)	Response (%)	Response (%)	Response (%)
GKi3_G1i1		12 (.96)	2 (.02)	10 (.01)	7 (.01)	5 (<.01)
GKi4_G1i2		15 (.66)	8 (.30)	20 (.02)	5 (.01)	3 (.01)
GKi5_G1i3_G2i1		5 (.62)	13 (.25)	9 (.06)	4 (.03)	6 (.03)
G1i4_G2i2		9 (.62)	43 (.18)	26 (.07)	8 (.06)	11 (.06)
G1i5_G2i3		6 (.48)	70 (.27)	60 (.13)	50 (.09)	0 (.02)
GKG1G2i6		Y (.99)	N (.01)	DNS (<.01)	UI (<.01)	–
GKG1G2i7		N (.99)	Y (.01)	DNS (<.01)	–	–
GKG1G2i8		Y (.64)	N (.35)	DNS (<.01)	UI (<.01)	–
GKG1G2i9		Y (.78)	N (.21)	DNS (<.01)	UI (<.01)	MR (<.01)
GKG1G2i10		N (.83)	Y (.17)	DNS (.01)	UI (<.01)	–
GKG1G2i11		Y (.73)	N (.26)	DNS (.01)	UI (<.01)	–
GKG1G2i12		Y (.81)	N (.18)	DNS (.01)	UI (<.01)	–
GKG1G2i13		Y (.57)	N (.42)	DNS (.01)	UI (<.01)	MR (<.01)
GKG1i14		29 (.75)	20 (.10)	31 (.06)	2 (.06)	40 (.03)
GKG1i15		26 (.76)	17 (.07)	20 (.06)	25 (.05)	6 (.05)
G1i16_G2i14		37 (.71)	30 (.10)	36 (.07)	46 (.06)	40 (.05)
GKG1i17		11 (.93)	10 (.02)	12 (.01)	9 (.01)	7 (.01)
GKG1i18_G2i15		14 (.85)	15 (.03)	13 (.03)	16 (.01)	12 (.01)
GKi20_G1i19_G2i16		5 (.80)	6 (.05)	19 (.04)	4 (.02)	7 (.02)
G1i20_G2i17		7 (.74)	6 (.06)	8 (.04)	25 (.03)	5 (.02)
G1i21_G2i18		8 (.77)	7 (.05)	18 (.03)	9 (.03)	DNS (.02)
GKi21_G1i22		35 (.81)	15 (.02)	30 (.02)	DNS (.02)	34 (.02)
G1i23_G2i19		99 (.54)	100 (.15)	105 (.05)	98 (.04)	101 (.02)
G1i24_G2i20		2 (.28)	18 (.09)	1 (.08)	10 (.04)	11 (.04)
GKi22_G1i25_G2i23		5 (.40)	9 (.35)	13 (.06)	4 (.03)	8 (.03)
G1i26_G2i24		1 (.38)	5 (.39)	9 (.06)	3 (.03)	DNS (.02)

Note. *n* = 1,821 valid grade 1 tests conducted. Items that were not answered were recorded as "DNS." Item responses that were unclear were recorded as "UI."

Table I.3. Proportion of Grade 2 Student Responses by Item

Item	Item description	Correct response Response (%)	Most frequent incorrect responses			
			Response (%)	Response (%)	Response (%)	Response (%)
GKi5_G1i3_G2i1		5 (.82)	13 (.12)	9 (.02)	4 (.02)	6 (.01)
G1i4_G2i2		9 (.74)	43 (.13)	11 (.07)	26 (.03)	8 (.02)
G1i5_G2i3		6 (.60)	70 (.17)	50 (.10)	60 (.10)	0 (.02)
G2i4		25 (.79)	65 (.11)	20 (.07)	45 (.02)	DNS (.01)
G2i5		140 (.56)	39 (.17)	120 (.13)	105 (.09)	31 (.04)
GKG1G2i6		Y (.99)	N (.01)	DNS (<.01)	—	—
GKG1G2i7		N (.99)	Y (.01)	DNS (<.01)	—	—
GKG1G2i8		Y (.66)	N (.34)	DNS (<.01)	UI (<.01)	—
GKG1G2i9		Y (.85)	N (.14)	DNS (<.01)	—	—
GKG1G2i10		N (.79)	Y (.21)	DNS (<.01)	UI (<.01)	—
GKG1G2i11		Y (.78)	N (.22)	DNS (<.01)	—	—
GKG1G2i12		Y (.86)	N (.13)	DNS (<.01)	UI (<.01)	MR (<.01)
GKG1G2i13		Y (.62)	N (.37)	DNS (.01)	—	—
G1i16_G2i14		37 (.84)	30 (.07)	36 (.04)	40 (.02)	46 (.02)
GKG1i18_G2i15		14 (.92)	2 (.01)	15 (.01)	13 (.01)	12 (.01)
GKi20_G1i19		5 (.91)	19 (.02)	6 (.02)	4 (.01)	DNS (.01)
G1i20_G2i17		7 (.85)	6 (.03)	8 (.03)	25 (.02)	5 (.01)
G1i21_G2i18		8 (.90)	7 (.03)	18 (.01)	9 (.01)	DNS (.01)
G1i23_G2i19		99 (.72)	100 (.08)	98 (.03)	101 (.03)	109 (.03)
G1i24_G2i20		2 (.56)	18 (.13)	1 (.05)	12 (.05)	11 (.02)
G2i21		6 (.33)	194 (.14)	16 (.07)	116 (.04)	106 (.04)
G2i22		686 (.54)	676 (.04)	576 (.03)	DNS (.03)	112 (.02)
GKi22_G1i25_G2i23		5 (.51)	9 (.33)	13 (.06)	DNS (.02)	2 (.01)
G1i26_G2i24		1 (.49)	5 (.37)	9 (.06)	DNS (.02)	3 (.01)
G2i25		101 (.37)	100 (.21)	DNS (.03)	175 (.03)	99 (.02)

Note. n = 1,764 valid grade 2 tests conducted. Items that were not answered were recorded as "DNS." Item responses that were unclear were recorded as "UI."