

# Elementary Mathematics Student Assessment

Measuring the Performance of Grade K, 1, and 2  
Students in Counting, Word Problems, and  
Computation in Fall 2015

Robert C. Schoen  
Daniel Anderson  
Zachary Champagne  
Charity Bauduin

MAY 2017

Research Report No. 2017-20

SECURE VERSION

The research and development reported here was supported by the Florida Department of Education, through Award Numbers 371-2355B-5C001, 371-2356B-6C001, and 371-2357B-7C004 to Florida State University. The opinions expressed are those of the authors and do not represent views of the Florida Department of Education.

Suggested citation: 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. doi:10.17125/fsu.1522170756.

Copyright 2017, Florida State University. All rights reserved. Requests for permission to use this test should be directed to Robert Schoen, [rschoen@lsi.fsu.edu](mailto:rschoen@lsi.fsu.edu), FSU Learning Systems Institute, 4600 University Center C, Tallahassee, FL, 32306.

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](mailto:rschoen@lsi.fsu.edu)).

# **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

**Robert C. Schoen<sup>1</sup>**

**Daniel Anderson<sup>2</sup>**

**Zachary Champagne<sup>1</sup>**

**Charity Bauduin<sup>1</sup>**

May 2017

<sup>1</sup>Florida State University

<sup>2</sup>University of Oregon

Florida Center for Research in Science, Technology, Engineering, and Mathematics (FCR-STEM)  
Learning Systems Institute  
Florida State University  
Tallahassee, FL 32306  
(850) 644-2570

## Acknowledgements

The successful development and implementation of this assessment involved many experts in mathematics education and many more students. Some of the key people involved with the 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 coordinated 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. Zachary Champagne assisted with the development and production of the test. 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.

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, Jiaqui Lu, Senai Tazaz, and Shelby McCrackin served as data-entry personnel. Claire Riddell, Mark McClure, and Monica Hurdal served as reviewers of the test items, response options, and scoring. Anne 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 for the students, parents, principals, district leaders, and teachers who agreed to participate in the study.

## Table of Contents

Acknowledgements .....	iv
Executive Summary .....	xii
Purpose .....	xii
Content .....	xii
Sample and Setting .....	xii
Test Specifications and Administration.....	xiii
Scoring.....	xiii
Reliability.....	xiii
Predictive Validity .....	xiii
Summary .....	xiii
1. Introduction and Overview .....	1
1.1. Test Overview .....	2
1.1.1. Counting Section .....	3
1.1.2. Word Problems Section.....	3
1.1.4. Computation Section.....	4
1.1.5. Detailed Test Blueprint.....	5
1.2. Test Administration.....	7
1.3. Description of the Sample and Setting.....	7
2. Test Development, Scoring, and Data Entry Procedures .....	9
2.1. Content.....	9
2.3. Instrument Development Process .....	9
2.4. Test Design and Assembly .....	10
2.5. Test Production .....	11
2.6. Test Administration for the Fall 2015 Field Test .....	11
2.7. Data Entry and Verification Procedures .....	12
2.8. Item-scoring Procedures .....	12
3. Data Analytic Procedures .....	14
3.1. Initial Screening According to Classical Test Theory .....	14
3.1.1. Classical item difficulty .....	14

3.1.2. Classical item discrimination.....	14
3.1.3. Item/raw score plots .....	15
3.2. Exploratory Factor Analysis .....	15
3.3. Specification of Models on the Basis of Item-response Theory .....	17
3.4. Vertical Linking.....	18
3.5. Predictive Validity .....	18
4. Results.....	19
4.1. Initial Screening of Items.....	19
4.2. Item Response Theory Models .....	22
4.3. Reliability .....	25
4.4. Predictive Validity .....	29
5. Discussion and Reflection .....	31
References .....	32

## List of Appendices

Appendix A. Grade K Test .....	34
Appendix B. Grade 1 Test.....	50
Appendix C. Grade 2 Test.....	68
Appendix D. Grade K Administration Guide.....	85
Appendix E. Grade 1 Administration Guide .....	98
Appendix F. Grade 2 Administration Guide .....	112
Appendix G. Scoring Key .....	126
Appendix H. Results of Initial Screening .....	129
H.1 Item-level Statistics.....	129
H.2 Spaghetti Plots.....	131
Appendix I. Most Common Incorrect Responses for Each Item .....	134

## List of Tables

Table 1.1. Final Blueprint for the Fall 2015 K–2 EMSA Test.....	2
Table 1.2. Items in the Counting Section.....	3
Table 1.3. Items in the Word Problems Section.....	4
Table 1.4. Items in the Computation Section.....	5
Table 1.5. Detailed Test Blueprint for the Fall 2015 K–2 EMSA.....	6
Table 1.6. Demographic Characteristics of the Students in the Fall 2015 Field-test of the K–2 EMSA Tests. .....	8
Table 2.1. Number of Times the Correct Answer is in Each Position.....	11
Table 3.1. Number of Factors Suggested by the MAP and VSS Tests.....	15
Table 4.1. Classical Test Theory-based Item Statistics for Items Removed from Scale during Screening Process.....	19
Table 4.2. Distribution of Item Difficulties and Discrimination Point Estimates for Items Used in the Final Scales.....	20
Table 4.3. Grade K Vertical Scale IRT Estimates.....	22
Table 4.4. Grade 1 Vertical Scale IRT Estimates.....	23
Table 4.5. Grade 2 Vertical Scale IRT Estimates.....	24
Table 4.6. Scaling Coefficients Used to Transform the Within-Grade Scales to a Common, Vertical Scale. .....	24
Table 4.7. Sample Descriptives for the Ability Estimates Generated by the Fall 2015 K–2 EMSA and Spring 2016 K–2 EMSA Tests, Split by Grade Level (Students with both Fall and Spring Scores Only).....	30
Table G.1. Grade K Scoring Key.....	126
Table G.2. Grade 1 Scoring Key.....	127
Table G.3. Grade 2 Scoring Key.....	128
Table H.1. Item Statistics for the Grade K Test Based on the Grade K Sample (n = 986).....	129
Table H.2. Item Statistics for the Grade 1 Test Based on the Grade 1 Sample (n = 1,763).....	130
Table H.3. Item Statistics for the Grade 2 Test Based on the Grade 2 Sample (n = 1,737).....	131
Table I.1. Proportion of Grade K Student Responses by Item.....	134
Table I.2. Proportion of Grade 1 Student Responses by Item.....	135
Table I.3. Proportion of Grade 2 Student Responses by Item.....	136



## List of Figures

Figure 2.1. One of the images used in place of page numbers. ....	10
Figure 3.1. Parallel analysis scree plot for the grade K test. ....	16
Figure 3.2. Parallel analysis scree plot for the grade 1 test. ....	16
Figure 3.3. Parallel analysis scree plot for the grade 2 test. ....	17
Figure 4.1. Distribution of the number of items answered correctly in the final, 11-item scale administered to the grade K sample (n = 986). ....	21
Figure 4.2. Distribution of the number of items answered correctly in the final, 20-item scale administered to the grade 1 sample (n = 1,763). ....	21
Figure 4.3. Distribution of the number of items answered correctly in the final, 19-item scale administered to the grade 2 sample (n = 1,737). ....	22
Figure 4.4. Test characteristic curves for grades K, 1, and 2 after vertical equating. ....	25
Figure 4.5. Test information functions for Grades K, 1, and 2. ....	26
Figure 4.6. Grade K item-person plot. ....	27
Figure 4.7. Grade 1 item-person plot. ....	28
Figure 4.8. Grade 2 item-person plot. ....	29
Figure H.1. Grade K spaghetti plot. ....	132
Figure H.2. Grade 1 spaghetti plot. ....	132
Figure H.3. Grade 2 spaghetti plot. ....	133

## List of Equations

Equation 1. Two-parameter logistic (2PL) item response theory (IRT) model (1)..... 17

## List of Abbreviations

2PL .....	Two-parameter logistic
CCSS-M .....	Common Core State Standards for Mathematics
CDU .....	Compare Difference Unknown
CGI .....	Cognitively Guided Instruction
CTT .....	Classical Test Theory
ELL .....	English Language Learner
EMSA .....	Elementary Mathematics Student Assessment
FA .....	Factor Analysis
IRT .....	Item Response Theory
JCU .....	Join Change Unknown
JRU .....	Join Result Unknown
MAP .....	Minimal Average Partial
MD .....	Measurement Division
MG .....	Multiplication Grouping
MPAC .....	Mathematics Performance and Cognition
OMR .....	Object Mark Recognition
PD .....	Partitive Division
PPU .....	Part-Part-Whole Part Unknown
SRU .....	Separate Result Unknown
SWD .....	Students With Disabilities
VSS .....	Very Simple Structure

## Executive Summary

This report describes an assessment instrument called the Elementary Mathematics Student Assessment: Measuring the Performance of Grade K, 1, and 2 Students in Counting, Word Problems, and Computation in Fall 2015. In this report, we will refer to the test as the Fall 2015 K–2 EMSA.

The Fall 2015 K–2 EMSA measures students’ ability to solve problems involving number and operations and is designed to serve as a mathematics achievement test administered to students at the beginning of the school year at the early elementary level. The Fall 2015 K–2 EMSA has three major sections: Counting, Word Problems, and Computation.

### Purpose

The intended use of the Fall 2015 EMSA test was to serve as a baseline measure of student achievement for use as a covariate in a randomized controlled trial evaluating the impact of a teacher professional development program called Cognitively Guided Instruction (CGI) on student learning. The purpose of the current report is to create a reference document that describes the content of the test, the development process, and the process we used to create the final scale. The current report therefore focuses on the content of the test, administration protocol, scoring procedures, and psychometric properties for the achievement focus of the Fall 2015 K–2 EMSA.

### 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) and the Mathematics Florida Standards (Florida Department of Education, 2014; NGACBP & CCSSO, 2010). In a few instances, the content of the test extends beyond the CCSS-M for the given grade level. For example, the grade 1 test includes a word problem involving a *multiplication-grouping* situation. Although this problem type is not specifically referenced in the CCSS-M for grade 1, the item has been used in empirical research studies for grade 1 students. In addition, the grade 2 test includes both a *partitive* and a *measurement division* word problem (Carpenter et al., 2015). Again, these problem types are not specifically referenced in the K–2 CCSS-M, but they are used on this assessment.

The Fall 2015 K–2 EMSA was designed to measure student achievement on types of problems that tend to be more difficult for students, so as to increase the ability of the test to discriminate among different levels of knowledge and understanding. For example, multidigit subtraction problems involved regrouping (i.e., borrowing) at least once. These types of numbers in subtraction problems are more likely to produce student errors resulting from limited understanding than are subtraction problems that do not involve regrouping. The problems in other sections also included more complex types and therefore more places for students to make errors. Analysis of the resulting data indicates that the test difficulty may have been too high, especially for grade K students.

### Sample and Setting

The Fall 2015 K–2 EMSA tests were administered to a total of 4,486 participating grade K, 1, and 2 students in 67 schools located in 10 public school districts in Florida during fall 2015.<sup>1</sup> The sample

---

<sup>1</sup> The Administration Guides provided in Appendices D, E, and F show 13 school districts. Some of those districts only had grades 3–5 teachers participating and are not part of this report. However, the beginning pages were the same and were used in all grades K–5.

included 968 grade K, 1,763 grade 1, and 1,737 grade 2 students. The school districts were implementing a curriculum based on the Mathematics Florida Standards (Florida Department of Education, 2014), which are very similar to the Common Core State Standards for Mathematics (CCSS-M; NGACBP & CCSSO, 2010).

## Test Specifications and Administration

The Fall 2015 K–2 EMSA includes selected-response test items at each grade level. The students are 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 responses students provided in previous administrations of items when items were presented in a constructed response format. The response options are presented horizontally, centered on the page, and the five response options are sequenced left to right with numbers from least to greatest. The grade K test includes 13 items, the grade 1 test 21 items, and the grade 2 test 20 items.

## Scoring

Student forms for the Fall 2015 K-2 EMSA were designed to be compatible with optic mark recognition (OMR) software in an attempt to increase efficiency of the data-entry process (Remark Office OMR 2014, Service Pack 4). Research assistants scanned the student forms into the OMR program, which read and recorded student responses to each item. To ensure the accuracy of the data recorded, each page of the form is identified with a unique barcode. This barcode is used by the OMR software to identify the grade level of each form and to ensure every page is correctly scanned for each test form. As an additional step to verify the accuracy of the scanned data, research assistants entered a 10% sample of student tests into a FileMaker Pro database (FileMaker Pro, Version 14.1). These separate records were found to have 99% agreement on scored responses.

## Reliability

Analysis of test information functions and item-person plots indicated that the test had adequate reliability at grades 1 and 2 for its intended purpose. The reliability at grade K did not exceed the desired .80 threshold. Additional development of the grade K test to decrease the average item difficulty is recommended as a strategy for improving reliability at grade K.

## Predictive Validity

The Fall 2015 K–2 EMSA test scores explained 52% of the variance in the (nonequated) Spring 2016 K–2 EMSA test scores, suggesting that the Fall 2015 K–2 EMSA test was reasonably well-suited for its intended use as a baseline covariate for student achievement in the larger study.

## Summary

The content of the Fall 2015 K–2 EMSA test aligns with the grade-level expectations in the CCSS-M and the Mathematics Florida Standards in the area of number and operations. 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 reliability of the tests were adequately high at grades 1 and 2, but the reliability did not meet the .80 threshold with the grade K sample, probably because the test was too difficult for beginning-of-year grade K students. Better alignment between the overall difficulty of the test and student abilities in grade K may result in increased reliability. The Fall 2015 K–2 EMSA

test scores explained more than 50% of the variance in student test scores as measured in Spring 2016. Overall, the test appears to be reasonably well-suited for its intended purpose.

## 1. Introduction and Overview

The Fall 2015 K–2 EMSA was the result of an iterative process of development and feedback from a variety of experts. This test built on our work in the development and implementation of the fall 2013 and fall 2014 EMSA tests (Schoen, LaVenía, Bauduin, & Farina, 2016a; 2016b) and the spring 2014 and spring 2015 Mathematics Performance and Cognition (MPAC) Interviews (Schoen, LaVenía, Champagne, & Farina, 2016; Schoen, LaVenía, Champagne, Farina, & Tazaz, 2016). The Fall 2015 K–2 EMSA has three major sections: Counting, Word Problems, and Computation.

The Fall 2015 K–2 EMSA was designed to serve as a mathematics achievement test administered to students at the beginning of the school year. It was designed to measure students' ability to solve problems involving number and operations. It did not measure other domains of mathematics knowledge, such as geometry, measurement, probability, or data analysis. The intended use of the fall 2015 EMSA test was to serve as a pretest measure of student achievement that would be used as a covariate in a randomized controlled trial evaluating the impact of a teacher professional-development program called Cognitively Guided Instruction (CGI) on student learning.

The Fall 2015 K–2 EMSA consisted of three test forms, one for each of the three grade levels. These tests were used to create a vertically scaled score, by means of item-response theory, that is directly comparable across grades. The vertically scaled score increases statistical power in the randomized controlled 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 are designed to be administered in a whole-group setting in a paper-pencil format. Test administrators are given an administration guide explaining how to administer the tests, along with a script to use while administering them. Questions are read aloud to students, and students shade bubbles to indicate their responses to multiple-choice items. Test administrators are encouraged to allow students to use manipulatives in accordance with their typical classroom practice.

The current report focuses on the content, administration protocol, scoring procedures, and psychometric properties for the Fall 2015 K–2 EMSA test. 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 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 Fall 2015 K–2 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 these items 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 Fall 2015 K–2 EMSA Test*

Section	Number of items		
	Grade K	Grade 1	Grade 2
Counting	5	5	3
Word Problems	3	6	7
Computation	4	10	10
Total	13	21	20

By design, at least three items were identical within each section on test forms at adjacent grade levels, 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 booklet provided to them. 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 Counting and Word Problems 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 problem was read by the administrator, the students were instructed to "fill in the bubble under the correct answer." Students were provided the time necessary to solve each problem, and test administrators were instructed to wait until all students were finished before moving to the next problem.

For the Computation section, test administrators were instructed to continue reading each problem to the grade K students, many of whom were experiencing school-based achievement testing for the first time. The grade 1 and grade 2 students completed this section independently.

The testing conditions for the Fall 2015 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 Fall 2015 K–2 EMSA if they were ordinarily permitted to do so in that particular classroom.



The Fall 2015 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. Counting Section

Table 1.2 provides an overview of the Counting items by grade level. The anchor set for grades K–1 include three items. The anchor set for grades 1–2 also includes three. One item in this section ( ) was included at all three grade levels, to create a set of anchor items among the three grade levels.

Table 1.2. Items in the Counting Section

Variable name	Grade K	Grade 1	Grade 2
GKi2			
GKi3			
GKi4_G1i1			
GKi5			
GKi6_G1i2			
GKi7_G1i3_G2i1			
G1i4_G2i2			
G1i5G2i3			

For each grade level, all the items in this section were read aloud twice to the students. This was done to lessen the effect of reading ability, listening comprehension, or working memory on test scores in attempt to focus the test on assessment of students’ mathematics achievement.

### 1.1.2. 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). The problems are sequenced, in general, from easier to more difficult.

Table 1.3. Items in the Word Problems Section

Variable name	Grade K	Grade 1	Grade 2
GKi8_G1i6			
GKi9_G1i7			
GKi10_G1i8_G2i5			
G1i9_G2i6			
G1i10_G2i4			
G1i11_G2i7			
G2i8			
G2i9			
G2i10			

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

Table 1.3 shows the types of word problems included in the Word Problems section of the Fall 2015 K–2 EMSA test at each grade level. The 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 given numbers in the problem.

As indicated in Table 1.3, the grades K–1 anchor set included three identical items, as did the grades 1–2 anchor set. One item was included in identical form at all three grade levels in this section of the test.

The two *division* word problems and the *Multiplication-Grouping* problem were beyond the scope of the content of the CCSS-M at grades 1 and 2. We included them, because abundant empirical evidence demonstrates that students at these grade levels can solve such 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—consistent with *Multiplication-Grouping* and *Measurement-Division* problems (Carpenter et al., 2015).

As for the previous section, all the items in this section at each grade level were read aloud twice to the students as per the administration instructions.

### 1.1.4. Computation Section

Table 1.4 provides an overview of the items in the Computation section of the Fall 2015 K–2 EMSA test at each grade level. In this section, the grades K–1 anchor set included four items; the grades 1–2 anchor set included six items. Three of the items in this section were identical across all three grade levels.

Table 1.4. Items in the Computation Section

Variable name	Grade K	Grade 1	Grade 2
GKG2i11_G1i12			
GKG2i12_G1i13			
GKi13_G1i15			
GKi14_G1i16_G2i13			
G1i14_G2i16			
G1i17_G2i14			
G1i18			
G1i19			
G1i20			
G1i21_G2i17			
G2i15			
G2i18			
G2i19			
G2i20			

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.

For this section, the items were expected to be administered differently at different grade levels. Test administrators were instructed to continue to read each problem aloud twice to grade K students. The additional instructions acknowledge that many of these children will have had little or no experience taking similar tests.

Unlike the grade K students, grade 1 and 2 students completed the Computation section independently. 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. They were encouraged to look closely at the symbol to decide whether each problem involved addition or subtraction.

For grade K students, the addition or subtraction problems were read aloud, as per the administration script, in two different ways. For example, was read aloud as, “ ? ? Fill in the bubble under the answer you think is correct. Again: ? ? Fill in the bubble under the answer you think is correct.” The problem was read aloud as, “ ? ? Fill in the bubble under the answer you think is correct. Again: ? ? Fill in the bubble under the answer you think is correct.” This language is consistent with that of other large-scale, standardized tests (e.g., Dunbar et al., 2008).

### 1.1.5. Detailed Test Blueprint

Table 1.5 provides a detailed blueprint showing the items in each of the three sections of the test (i.e., Counting, Word Problems, 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 1.5. Detailed Test Blueprint for the Fall 2015 K–2 EMSA

Item description	Variable names		
	Grade K	Grade 1	Grade 2
<i>Counting</i>			
	GKi2		
	GKi3		
	GKi4_G1i1	<del>GKi4_G1i1</del>	
	GKi5		
	GKi6_G1i2	GKi6_G1i2	
	GKi7_G1i3_G2i1	GKi7_G1i3_G2i1	GKi7_G1i3_G2i1
		G1i4_G2i2	G1i4_G2i2
		G1i5_G2i3	G1i5_G2i3
<i>Word Problems</i>			
	GKi8_G1i6	GKi8_G1i6	
	GKi9_G1i7	GKi9_G1i7	
	GKi10_G1i8_G2i5	GKi10_G1i8_G2i5	GKi10_G1i8_G2i5
		G1i9_G2i6	G1i9_G2i6
		G1i10_G2i4	G1i10_G2i4
		G1i11_G2i7	G1i11_G2i7
			G2i8
			G2i9
			G2i10
<i>Computation</i>			
	GKG2i11_G1i12	GKG2i11_G1i12	<del>GKG2i11_G1i12</del>
	GKG2i12_G1i13	GKG2i12_G1i13	GKG2i12_G1i13
	<del>GKi13_G1i15</del>	GKi13_G1i15	
	<del>GKi14_G1i16_G2i13</del>	GKi14_G1i16_G2i1	GKi14_G1i16_G2i13
		G1i14_G2i16	G1i14_G2i16
		G1i17_G2i14	G1i17_G2i14
		G1i18	
		G1i19	
		G1i20	
		G1i21_G2i17	G1i21_G2i17
			G2i15
			G2i18
			G2i19
			G2i20
Items on Test Form	13	21	20
Items in Final Scale	11	20	19

Note. The four items in strikethrough font were on the test form but were removed from the final scale at those respective grade levels as a result of poor item statistics.

## 1.2. Test Administration

Teachers were given detailed instructions on how to administer the test (which appear in Appendices D, E, and F), including a script to use during administration.

Teachers 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 two sections of the test, teachers were instructed to read the problems aloud to students—in their entirety—to reduce the effect of reading ability on students' mathematics performance. They 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 envelopes to the front office for project personnel to pick up during a window of time outlined in the administration instructions.

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

## 1.3. Description of the Sample 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 CGI. Tests forms were delivered to schools by project staff during the week of preplanning (i.e., the week before students return to school for the year). In the field tests reported in the present report, the students' classroom teachers administered the tests during the first two weeks of the school year in all but five classrooms.

There are a total of 4,468 students in the analytic sample, with 968 students representing grade K, 1,763 representing grade 1, and 1,737 representing grade 2. These students represent 266 classrooms in 10 Florida public school districts<sup>2</sup>. Table 1.6 provides descriptive statistics for the data we have at the time of this report.

---

<sup>2</sup> The Administration Guides provided in Appendices D, E, and F show 13 school districts. Some of those districts only had grades 3–5 teachers participating and are not part of this report. However, the beginning pages were the same and were used in all grades K–5.

Table 1.6. Demographic Characteristics of the Students in the Fall 2015 Field-test of the K–2 EMSA Tests

Student characteristic	Number (proportion of sample or subsample)			
	Grade K ( <i>n</i> = 986)	Grade 1 ( <i>n</i> = 1,763)	Grade 2 ( <i>n</i> = 1,737)	Overall sample ( <i>n</i> = 4,486)
<b>Gender</b>				
Male	150 (.15)	233 (.13)	231 (.13)	614 (.14)
Female	131 (.13)	220 (.12)	241 (.14)	592 (.13)
Unknown	705 (.72)	1,310 (.74)	1,265 (.73)	3,280 (.73)
<b>Language</b>				
ELL	24 (.02)	38 (.02)	21 (.01)	83 (.02)
Non-ELL	257 (.26)	415 (.24)	448 (.26)	1,120 (.25)
Unknown	705 (.72)	1,310 (.74)	1,268 (.73)	3,283 (.73)
<b>Exceptionality</b>				
SWD	12 (.01)	50 (.03)	39 (.02)	101 (.02)
Non-SWD	269 (.27)	403 (.23)	433 (.25)	1,105 (.25)
Gifted	7 (.01)	14 (.01)	42 (.02)	63 (.01)
Nongifted	274 (.28)	439 (.25)	430 (.25)	1,143 (.25)
Unknown	705 (.72)	1,310 (.74)	1,265 (.73)	3,280 (.73)
<b>Race</b>				
White	81 (.08)	118 (.07)	136 (.08)	335 (.07)
Black	29 (.03)	38 (.02)	36 (.02)	103 (.02)
Asian	5 (.01)	1 (<.01)	3 (<.01)	9 (<.01)
Other	20 (.02)	26 (.01)	28 (.02)	74 (.01)
Unknown	851 (.86)	1,580 (.90)	1,534 (.88)	3,965 (.88)
<b>Ethnicity</b>				
Hispanic	33 (.03)	31 (.02)	16 (.01)	80 (.02)
Non-Hispanic	135 (.14)	183 (.10)	203 (.12)	521 (.12)
Unknown	818 (.83)	1,549 (.88)	1,518 (.87)	3,885 (.86)

*Note.* ELL = English language learner. SWD = Students with disabilities. A large proportion of individual student demographic characteristics, such as ethnicity, exceptionality, or eligibility for free or reduced-price lunch, were not available at the time the report was written. Some of the percentages do not sum to 1.00 as a result of rounding.

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 and operations, but it includes some items designed to favor students who have a solid grasp of place-value concepts. The numbers used on the test are limited to positive integers between 1 and 100. Computation items presented symbolically involve applying either the addition or the subtraction operation on exactly two positive integers. Problems involving subtraction result in a difference with a positive, integer value. Word problems involve additive situations as well as grouping situations that could be solved by multiplication, division, addition, counting strategies, or direct place-value understanding (Carpenter et al., 1999, 2015).

Test design involved finding an optimal point at the intersection of three 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, (2) to produce a reasonably strong student-level test covariate for students' baseline mathematics abilities in the randomized-controlled trial, and (3) to minimize the testing burden on teachers and students.

### 2.2. Instrument Development Process

The development process for the Fall 2015 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., 2016), the 2015 MPAC Interview (Schoen et al., 2016), and the Fall 2013 and Fall 2014 EMSA test items (Schoen et al., 2016a, 2016b)
3. Review of the content of the CGI professional-development plan
4. Development of the first written draft of the test blueprint
5. Review of the draft blueprint by internal members of the evaluation team and external 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 informed the determination of the set of

response options in the selected-response format of the Fall 2015 K–2 EMSA tests. In general, the five most frequently provided responses were used as the five response options.

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

In general, the response options for the items in the Word Problems section included the two given numbers 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 provide a way for students to indicate their response, and the bar code used for scoring each item. Plenty of empty space is available on the page for students to draw or record their thoughts as necessary. The Computation section consists of items presented as open equations. Each problem is presented as a single equation involving either the addition or the subtraction operator and exactly two numerals. Each is presented in the standard (i.e.,  $a + b = c$ ,  $a - b = c$ ) form (Stigler et al., 1986; Schoen et al., in review) with an open box for the missing number. Students fill in the oval beneath the numeral to indicate their responses.

In the Counting and Word Problems sections, only one problem is displayed per page so that students will not record their answers in the wrong places or be overwhelmed by too much text on the page. For grades 1 and 2, Computation items are presented with multiple items split across two pages. In an effort to avoid confusion, a line is placed after each Computation item on the page. For grade K, the Computation section includes only one item per page. The grammar used in word problems was reviewed by people with expertise in teaching emergent bilingual students. Large (20-point) Calibri font was used on the Counting and Word Problems sections in the final version of the grade K test. Large (48-point) Calibri was used on the Computation section in the final version of the grade K test. Both grades 1 and 2 tests used 18-point Calibri on the Counting and Word Problems sections in the final version, and 36-point Calibri on the Computation section. Copies of the grades K, 1, and 2 tests are presented in Appendices A, B, and C, respectively.

Because beginning-of-year grade K and 1 students, in particular, may not yet be able to read Arabic numerals, pages were identified by a series of child-friendly images rather than page numbers. Figure 1.2 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.

Pages were also identified by barcodes printed at the bottom of each page. The barcodes were used as identifiers for the OMR software to ensure it was using the correct template for each page it was reading. The barcodes did not include letters or numerals.



Every item on the Fall 2015 K–2 EMSA tests was presented in a selected-response format. Five response options was presented horizontally across the page and included exactly one correct response for each item. The response options were always numerals and were ordered from least to greatest, from left to right. The students were directed to fill in the circles (which we call bubbles) below their answer choices. Bubbles are centered beneath the corresponding response option, and responses are centered horizontally across the page. During the test development, careful consideration was given to the frequency of the correct-response positions, as well as to the distribution of correct-response positions across each test form to make them approximately evenly distributed across the various positions. 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 is in Each Position*

Grade level	A	B	C	D	E
K	3	2	2	4	2
1	3	6	4	5	3
2	4	5	2	5	4

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 shading the bubble). The set of incorrect responses (distractors) consisted of the most frequently encountered incorrect student responses in open-ended versions of the items on the Fall 2013 and Fall 2014 grades 1–2 EMSA tests and the 2014 and 2015 MPAC interviews, as well as other sources. Response options in the word problem section also usually contained the two numbers in the problem, their sum, and their difference.

Test administrators are directed to read each math problem aloud to students in accordance with the administration script. In addition, they are asked to provide and allow students to use manipulatives, like counters or linking cubes, during the test. If students require testing accommodations resulting from IEP, ELL or 504 plans, then the test administrator is 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 are instructed to allow students adequate time to answer all of the questions.

## 2.4. Test Production

The tests were printed on 28-pound, white paper at Florida State University and distributed to the participating schools. Those for grade K, were printed single-sided to reduce confusion among beginning-of-year kindergarten students. For grades 1 and 2, they were printed double-sided. The heavy paper was used, because the optical scanner yields better scanning results with it than with the more economical 20-pound paper. Administration guides and consent forms were printed on 20-pound, white paper at Florida State University.

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 Fall 2015 Field Test

Each participating teacher was provided with a test packet containing

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

They were distributed to the main offices at school sites during the week of preplanning. These materials were then distributed to the participating teacher from the main office personnel or principal-appointed designee. Teachers were instructed to administer the tests during the first two weeks of school.

The Fall 2015 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. In addition, the administration guides included a student information sheet on the last page. Test administrators used this sheet to provide student and class information (e.g., student names, student ID numbers, testing accommodations provided) and returned it with the completed student tests. The final forms of the test administration guides for grades K, 1 and 2 are presented in Appendices D, E, and F, respectively.

Upon conclusion of administration, teachers were instructed to submit all testing materials (test administration guide, student test booklets, student information sheet, and parental consent forms) to their principals or designees. Teachers were asked to return only completed test booklets completed by those students with corresponding signed parental consent on the parental consent form. The principal or designee placed the testing materials in the main office at the front desk for pickup. Members of the project team picked up test materials during the first two weeks of September 2015.

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. Five 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**

Data from the grade K, 1, and 2 Fall 2015 K–2 EMSA were recorded by means of OMR software. Tests were scanned on Fujitsu high-volume scanners and read by Remark OMR software. Bar codes designated each page of the assessment and were also used as student identifiers. The page identifiers were used to ensure no pages were skipped or shuffled out of place, whereas the student identifiers ensured that data from each test was associated with the correct student. The tests were designed to be read by the software used. The spacing for all items and responses were designed according to specifications provided by Remark. The OMR software was programmed to read and record the page identifier as well as the item responses on the page. These identifiers were printed on each page of the tests as a bar code. Research assistants prepared each returned test form to ensure responses could be accurately read by the OMR software. Errant marks were removed, and any lightly shaded responses were filled in darker. As a test of accuracy, a 10% sample of the data were also manually entered into a FileMaker database by trained data-entry staff, and the responses were compared with the OMR recorded responses. These separate records were found to have 99% agreement on scored responses.

## 2.7. Item-scoring Procedures

Every item on the Fall 2015 K–2 EMSA test forms used a selected-response format. Tests were scored by means of the scoring key created and reviewed during the test-development process. The scoring key was used to transform the raw responses into a dichotomous (correct, incorrect) variable. The scoring guide can be found in Appendix G.

## 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 Fall 2015 EMSA were subjected to the following analyses:

1. Initial screening of items by means of classical test theory (CTT)
2. Exploratory Factor Analysis
3. Within-grade scaling according to a two-parameter logistic item-response theory (2PL-IRT) model
4. Equating of scales between grades to create the vertical scale using the Stocking-Lord method (Kolen & Brennan, 2014)
5. Examination of the ability of Fall 2015 K–2 EMSA scores to predict students' Spring 2016 K–2 EMSA scores<sup>3</sup>

Initial item screening with CTT was completed to identify items that might not be providing useful information about test-takers' abilities. 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 construct or for a multidimensional construct. As described in greater detail below, the results of the factor analyses supported an essentially unidimensional measure, and scaling proceeded accordingly. The following sections provide more detailed information about the analytic processes we used.

### 3.1. Initial Screening According to Classical Test Theory

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 cut points (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 cut points were not considered as strict rules. 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 was removed or not.

#### 3.1.1. Classical item difficulty

Each individual item on the Fall 2015 K–2 EMSA was scored dichotomously. For these items, the CTT 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 true. 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).

#### 3.1.2. Classical item discrimination

Items were considered to have adequate discrimination if high-ability students tended to answer correctly and low-ability students to answer incorrectly. Using a classical approach, the item

---

<sup>3</sup> Fall 2015 and Spring 2016 EMSA tests were not equated with one another

discrimination was assessed by examination of the relation between test-takers' performance on each individual item and their total raw score (total number of correct items). This correlation was calculated for each item on each test using R-statistical environment (R Core Team, 2017). 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 below 0.20 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. 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	1	7	7
1	2	1	2
2	1	1	2

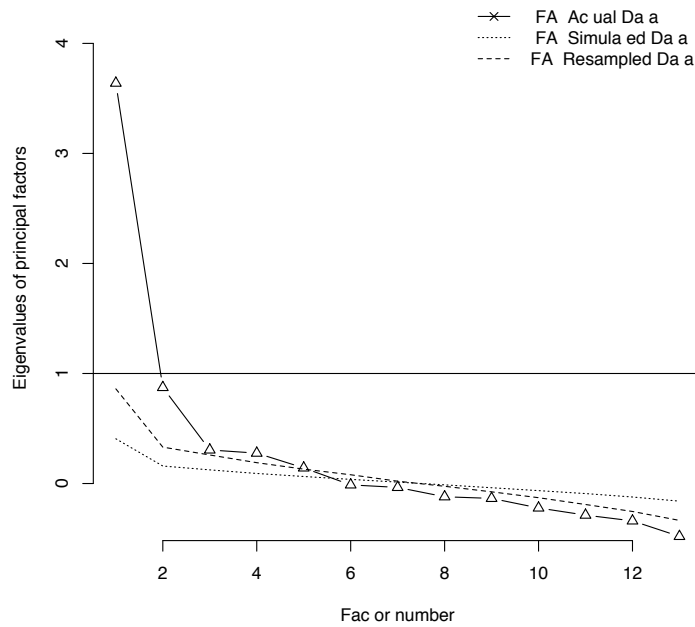


Figure 3.1. Parallel analysis scree plot for the grade K test.

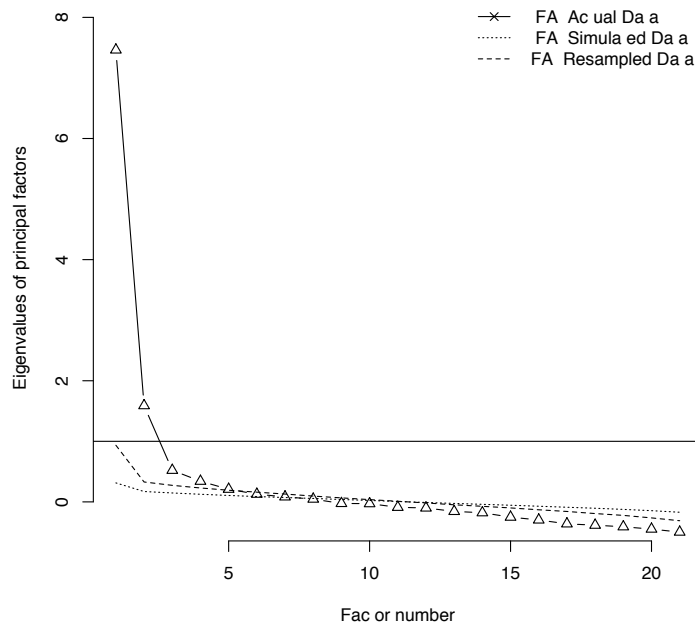


Figure 3.2. Parallel analysis scree plot for the grade 1 test.

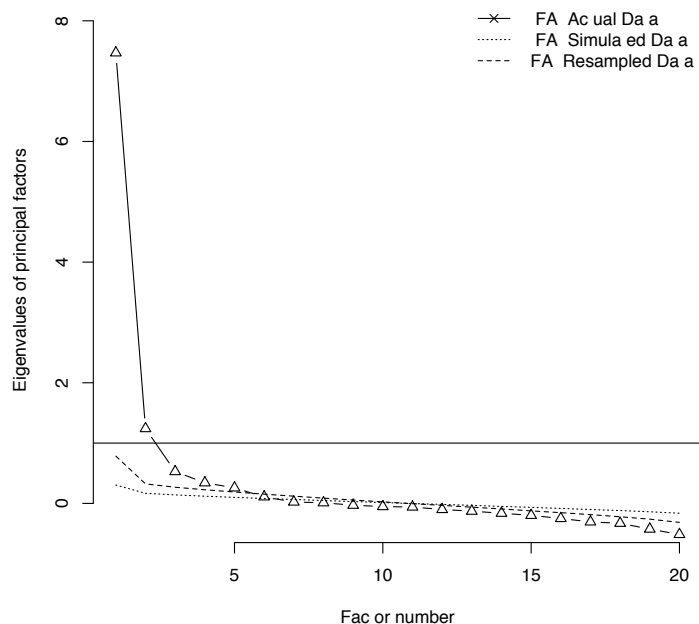


Figure 3.3. Parallel analysis scree plot for the grade 2 test.

Across all tests of the number of dimensions to extract, at least one always indicated a unidimensional structure. The scree plots also universally 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 unidimensional for practical applied purposes. Further, recent evidence from Anderson et al. (2017) suggests the 2PL IRT model is robust to mild deviations from unidimensionality. Given the collective evidence, and 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.

### 3.3. Specification of Models on the Basis of Item-response Theory

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

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

where  $\theta_j$  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,  $\theta_j$ , and the difficulty of the item  $b_i$ . Log odds are estimated as the ratio between the odds of a correct rather than 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  $\theta$  increases). Items with lower discrimination values are

weighted less in the estimation of theta 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.

These initial models served as an additional source of item screening; items with overly low or high discrimination estimates were evaluated by content experts for removal. Items that were overly difficult or easy were also marked for potential removal.

### 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).

### 3.5. Predictive Validity

The ability estimates generated with the Fall 2015 K–2 EMSA tests are designed to be used in a larger study involving a randomized controlled trial designed to estimate the effect of a teacher professional-development program on student achievement. The ability estimates will be used to test for baseline equivalence of the schools assigned to the treatment conditions and as a student achievement baseline covariate in multilevel models of analysis of covariance. On the basis of the students' scores on the test administered in spring 2016, we calculated how much of the variance was explained by those same students' scores on the Fall 2015 K–2 EMSA. This information can provide some evidence of external validity (Flake, Pek, & Hehman, 2017), and it is also useful in analysis of the statistical power in a given study.



These analyses involved first saving the scale scores from the final, vertically scaled scores for the grade K, 1, and 2 tests. Then, as manifest variables, the scale scores were merged into a file containing similar scores for the spring 2016 EMSA tests for grades K–2 (Schoen, Anderson, & Bauduin, 2017). We investigated evidence of predictive validity using a regression model in which the Fall 2015 K–2 EMSA scores predicted the Spring 2016 K–2 EMSA scores for each student in the sample with both fall and spring test scores. It should be noted that the fall 2015 and spring 2016 EMSA tests were not equated with one another.



## 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.

The initial calculations revealed that items GK13\_G1i15 and GK14\_G1i16\_G2i13 had a low  $p$ -values and point biserial correlations close to the .20 threshold. During review, content experts acknowledged that these subtraction items (  and  , respectively) may simply have been too difficult for students in the beginning of kindergarten (when students have very little knowledge of print symbols and are unlikely to have been introduced to the concept of the subtraction operation). On the basis of these results, both of these items were removed from the grade K IRT model, but both were retained on the grade 1 and 2 tests.

We also investigated item GK10\_G1i8\_G2i5, because only 9% of grade K students solved it correctly, but because this item had a point biserial correlation of .38 and fit nicely in the item/raw score plots with the other items, it was retained.

On the basis of the initial screening of items on the grade 1 test, we decided to drop GK14\_G1i1 from the grade 1 test before the subsequent IRT modeling. This counting item proved to be very easy for grade 1 students; 96% solved it correctly. It also had a low point-biserial correlation value of .18. Although this item was removed from the Grade 1 scale, it remained in the Grade K scale.

On the grade 2 test, we decided to drop GKG2i11\_G1i12 from the scale, because 93% of the grade-2 students solved it correctly. We kept this item to use in the grade K and 1 IRT-based models.

*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
GKi13_G1i15		K	.21 (.013)	.23
GKi14_G1i16_G2i13		K	.14 (.011)	.27
GKi4_G1i1		1	.96 (.005)	.18
GKG2i11_G1i12		2	.93 (.006)	.37

Table 4.2 shows the distribution of item difficulty and item discrimination for the items used at each of the three grade levels after items were removed based on the initial screening process.

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

Value	Number of items		
	Grade K	Grade 1	Grade 2
	<i>P-value</i>		
>.90	1	0	0
.80 – .89	1	1	4
.70 – .79	2	3	3
.60 – .69	0	1	3
.50 – .59	0	2	5
.40 – .49	0	2	2
.30 – .39	2	6	1
.20 – .29	3	5	1
.10 – .19	1	0	0
<.09	1	0	0
Mean	0.46	0.46	0.61
Median	0.36	0.38	0.62
Standard Deviation	0.29	0.19	0.19
	<i>Point-biserial correlation</i>		
.80 – 1.0	0	0	0
.60 – .79	1	5	3
.40 – .59	8	12	14
.20 – .39	2	3	2
0.0 – .20	0	0	0
Mean	0.47	0.50	0.50
Median	0.47	0.49	0.48
Standard Deviation	0.07	0.09	0.09

*Note.* Because all items were scored dichotomously, the p-value is the proportion of the sample judged as having provided a correct answer.

The distributions presented in Table 4.2 appear to show that the level of difficulty is lower for the grade 2 students than for the grades K and 1 students. Although the mean and median item difficulty are approximately the same for grades K and 1 items, the grade K item difficulty values have a bimodal distribution, and the grade 1 item difficulties are spread more evenly across the center of the distribution. Despite the clear differences in the distributions of item difficulties, the overall item discrimination appears to be roughly equivalent across all three grade levels.

The distribution of raw scores (i.e., number of items answered correctly) for the final set of items in the grade K, 1, and 2 tests are provided in Figures 4.1, 4.2, and 4.3, respectively. After the initial screening process, the total number of items on the grade K test was 11. The raw-score distribution for the grade K sample appears to be approximately symmetric, whereas the distributions appear to be slightly positively skewed for the grade 1 sample and slightly negatively skewed for the grade 2 sample. Approximately 1% of the grade K and 1 samples and about 2.5% of the grade 2 sample responded correctly to every item. Approximately 2.5% of the grade K students did not provide a correct response for any item, whereas all of the students in the grade 1 and 2 samples provided at least one correct response.

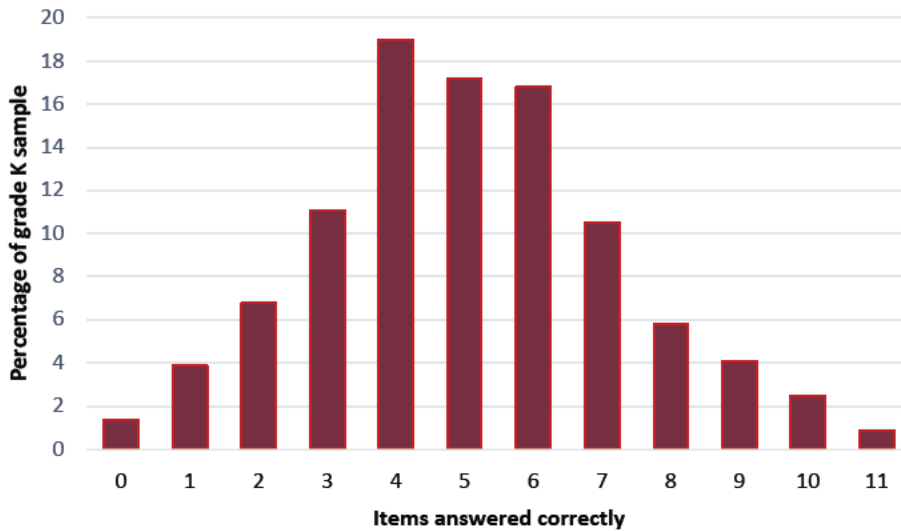


Figure 4.1. Distribution of the number of items answered correctly in the final, 11-item scale administered to the grade K sample ( $n = 986$ ).

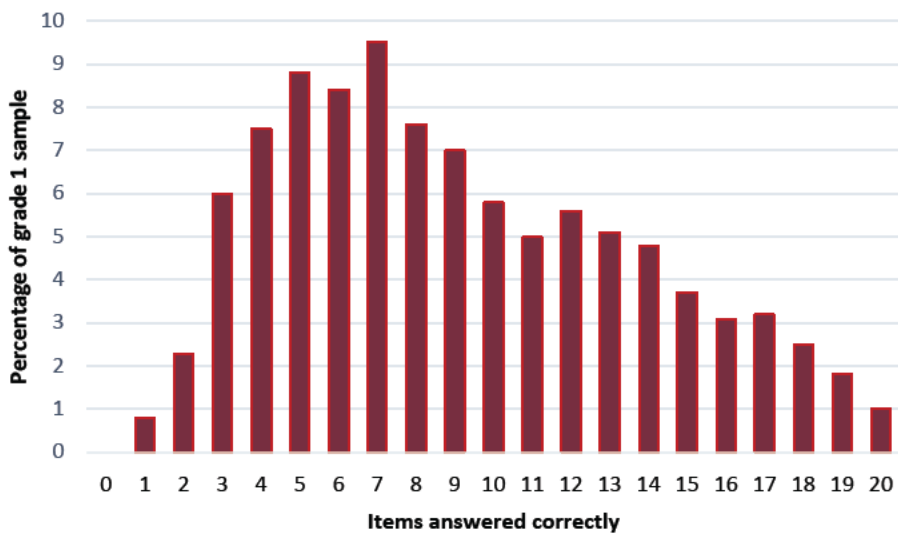


Figure 4.2. Distribution of the number of items answered correctly in the final, 20-item scale administered to the grade 1 sample ( $n = 1,763$ ).

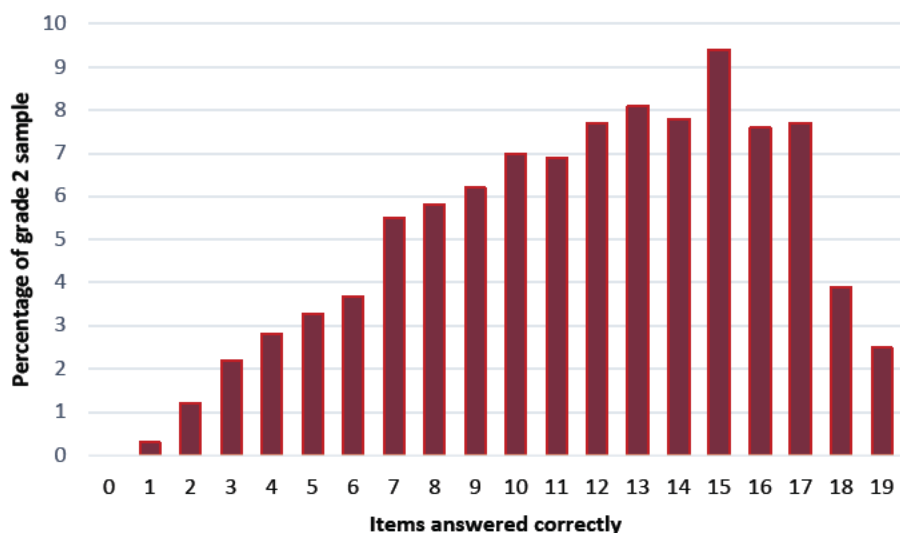


Figure 4.3. Distribution of the number of items answered correctly in the final, 19-item scale administered to the grade 2 sample ( $n = 1,737$ ).

## 4.2. Item Response Theory Models

The difficulty and discrimination estimates for the three grades resulting from a 2PL model are presented in Tables 4.3, 4.4, and 4.5. Note that these are the within-grade estimates and are therefore not comparable across grades. The mean item difficulties were -1.67, 0.15, and 1.01 for grades K–2, respectively. Item discriminations ranged from 0.74 to 2.64 in grade K, 0.59 to 2.88 in grade 1, and 0.43 to 1.93 in grade 2.

Table 4.3. Grade K Vertical Scale IRT Estimates

Item	Difficulty	Discrimination
GKG2i11_G1i12	-0.44	0.74
GKG2i12_G1i13	-0.52	1.24
GKi10_G1i8_G2i5	0.32	1.14
GKi2	-3.84	1.63
GKi3	-3.23	2.64
GKi4_G1i1	-2.82	2.25
GKi5	-2.99	1.68
GKi6_G1i2	-1.60	1.88
GKi7_G1i3_G2i1	-0.82	0.87
GKi8_G1i6	-1.56	1.04
GKi9_G1i7	-0.88	0.89

Table 4.4. Grade 1 Vertical Scale IRT Estimates

Item	Difficulty	Discrimination
G1i10_G2i4	0.59	0.84
G1i11_G2i7	1.53	0.93
G1i14_G2i16	0.75	0.87
G1i17_G2i14	0.48	2.55
G1i18	0.48	2.43
G1i19	0.39	2.36
G1i20	0.77	1.80
G1i21_G2i17	0.94	1.23
G1i4_G2i2	-0.65	0.59
G1i5_G2i3	1.73	0.75
G1i9_G2i6	1.24	1.05
GKG2i11_G1i12	-0.96	1.17
GKG2i12_G1i13	-0.30	0.90
GKi10_G1i8_G2i5	0.79	0.93
GKi13_G1i15	0.30	2.01
GKi14_G1i16_G2i13	0.26	2.88
GKi6_G1i2	-1.46	1.22
GKi7_G1i3_G2i1	-0.86	1.17
GKi8_G1i6	-1.97	0.72
GKi9_G1i7	-0.97	1.27

*Table 4.5. Grade 2 Vertical Scale IRT Estimates*

Item	Difficulty	Discrimination
G1i10_G2i4	0.92	1.36
G1i11_G2i7	1.34	1.93
G1i14_G2i16	0.41	0.86
G1i17_G2i14	0.26	1.25
G1i21_G2i17	0.75	0.75
G1i4_G2i2	-0.24	1.12
G1i5_G2i3	1.25	1.23
G1i9_G2i6	1.11	1.80
G2i10	2.01	1.24
G2i15	0.00	0.85
G2i18	1.51	0.80
G2i19	3.63	0.43
G2i20	3.57	0.71
G2i8	1.49	1.43
G2i9	1.67	1.52
GKG2i12_G1i13	-0.84	0.90
GKi10_G1i8_G2i5	1.28	1.39
GKi14_G1i16_G2i13	-0.30	0.99
GKi7_G1i3_G2i1	-0.55	1.37

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	0.93 (0.07)	-2.07 (0.09)
2	1	1.19 (0.05)	1.58 (0.06)

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 numbers of items differed for different grade level, affecting the heights of the curves in Figure 4.4. The curves indicate some separation between grade levels, a desirable feature of the scale, especially at these grade levels (where students change and learn very quickly). The vertical dashed lines can be interpreted as the estimated ability level associated with having a 50% chance of responding to approximately half the items correctly.

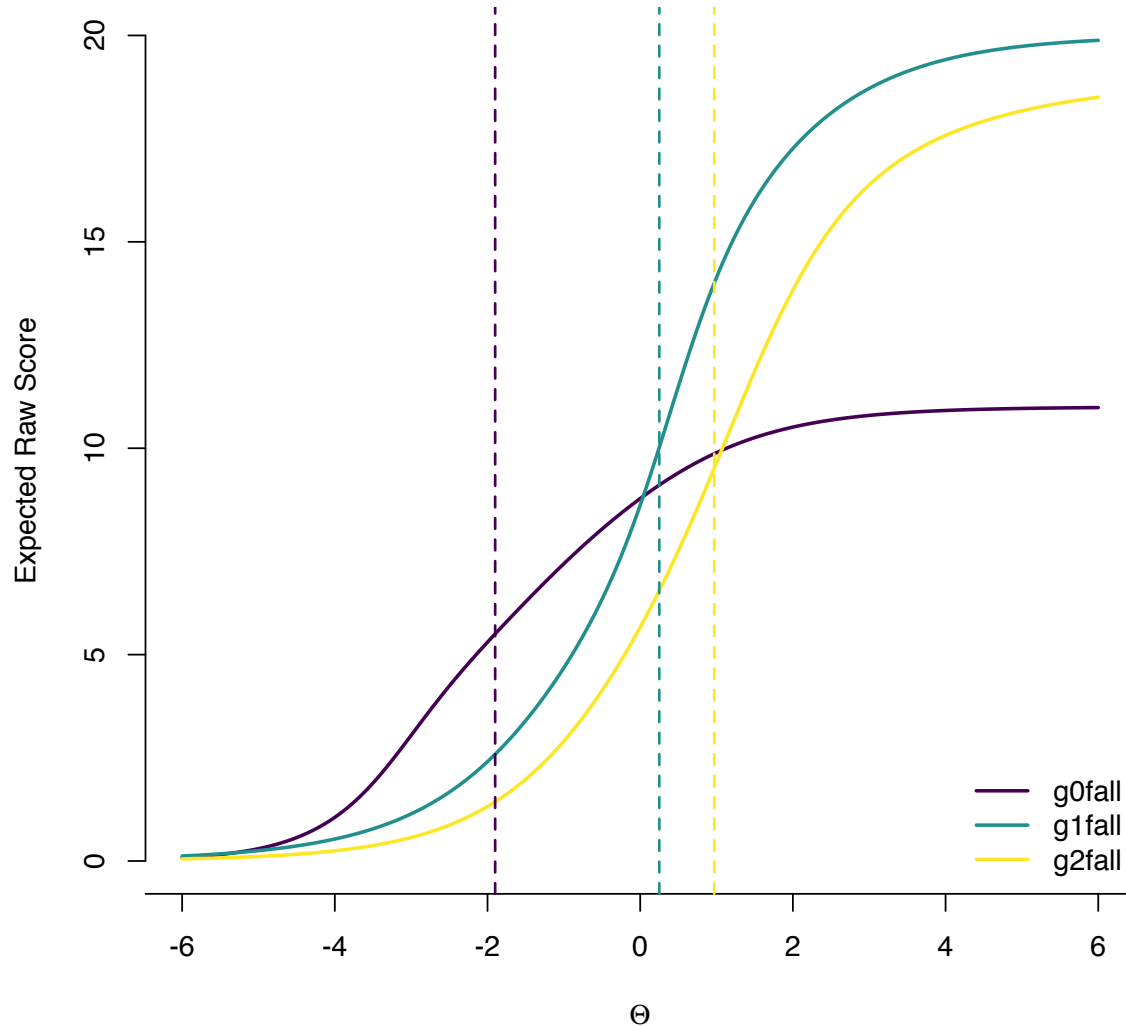


Figure 4.4. Test characteristic curves for grades K, 1, and 2 after vertical equating.

### 4.3. Reliability

Item response theory provides a conditional view of reliability, in which the reliability of the measure is viewed as depending upon the ability level of the respondent. This approach recognizes that reliability is not fixed but variable, depending on who is taking the test. Figure 4.5 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  $\theta$ , reliability is equivalent to a Cronbach’s alpha of 0.80 when information is equal to 5.0. Therefore, in the figure, 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  $\geq 0.80$ ). Notice that grade K (denoted in the plot as g0fall) did not eclipse reliability equivalent to 0.80 at any ability level.

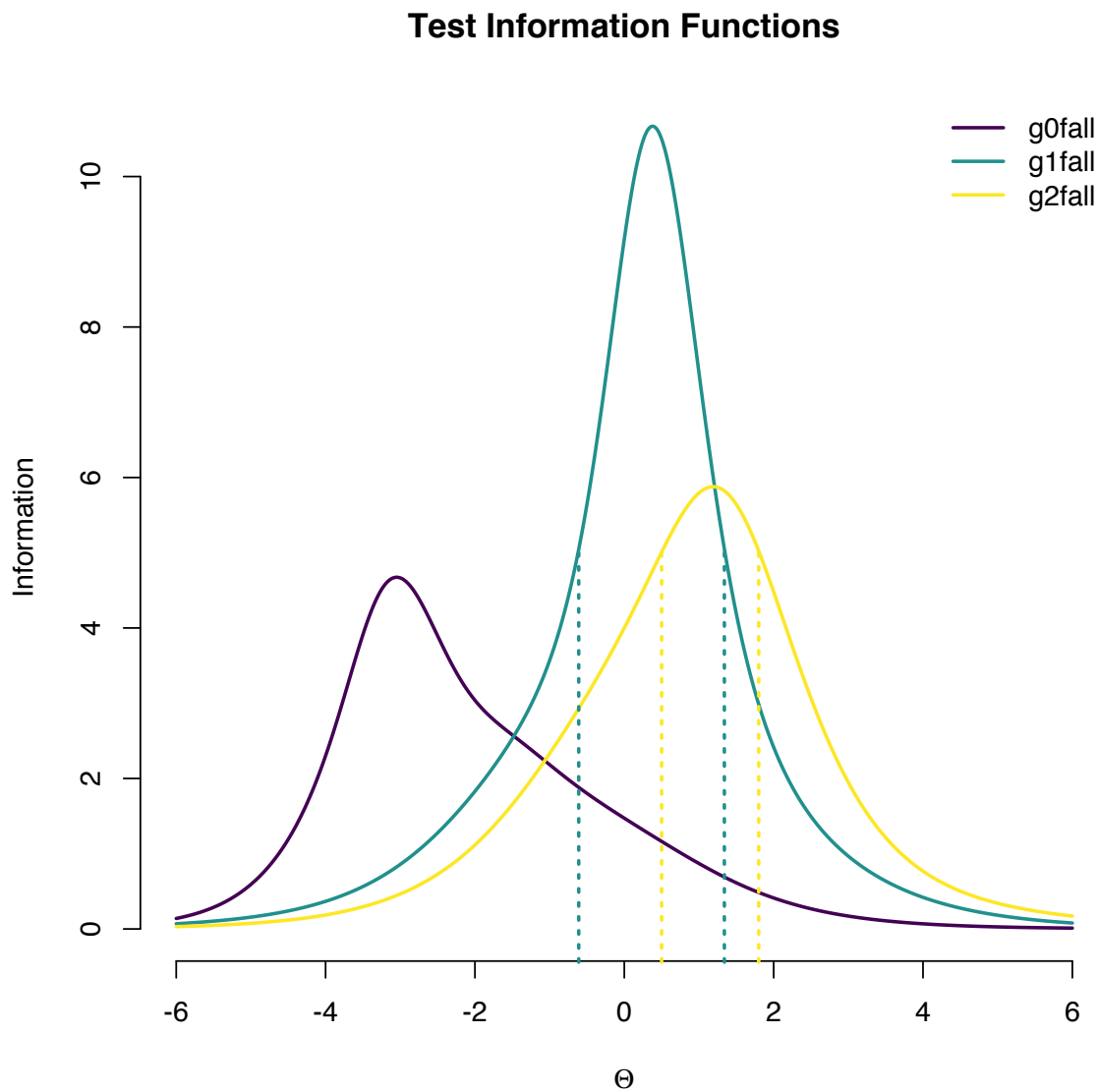


Figure 4.5. Test information functions for Grades K, 1, and 2.

Another way to understand the reliability of the instrument is to examine the match between the distribution 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 was somewhat bimodal, but the higher peak was at a higher estimated difficulty level than the peak for the person estimates, implying that a fair number of the items were more difficult than the median person ability. Generally, however, item difficulties matches student abilities well. At grade 1, the peak of the item difficulty distribution was again at a higher level than the person distribution. In these plots, the ranges in which reliability was equivalent to 0.80 or above are also displayed (as discussed above, no such range existed in grade K). The vertical blue 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. The same plot is displayed for grade 2. Note that at grade 2 the median item difficulty was less than the median person ability, reversing what was observed at grades K and 1.

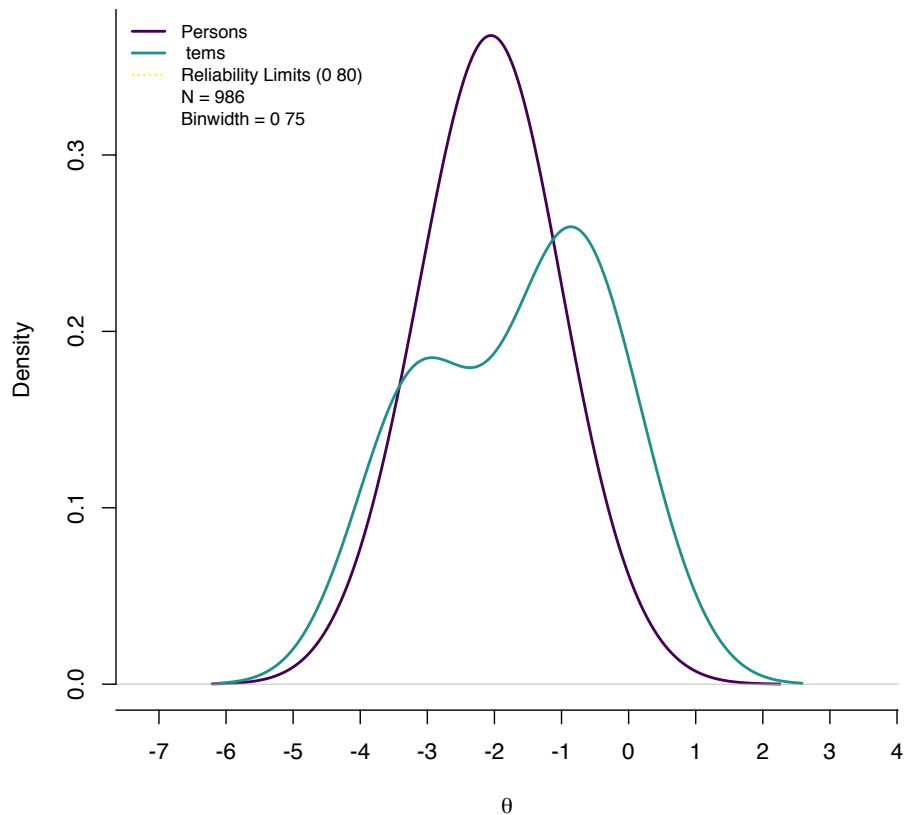


Figure 4.6. Grade K item-person plot.

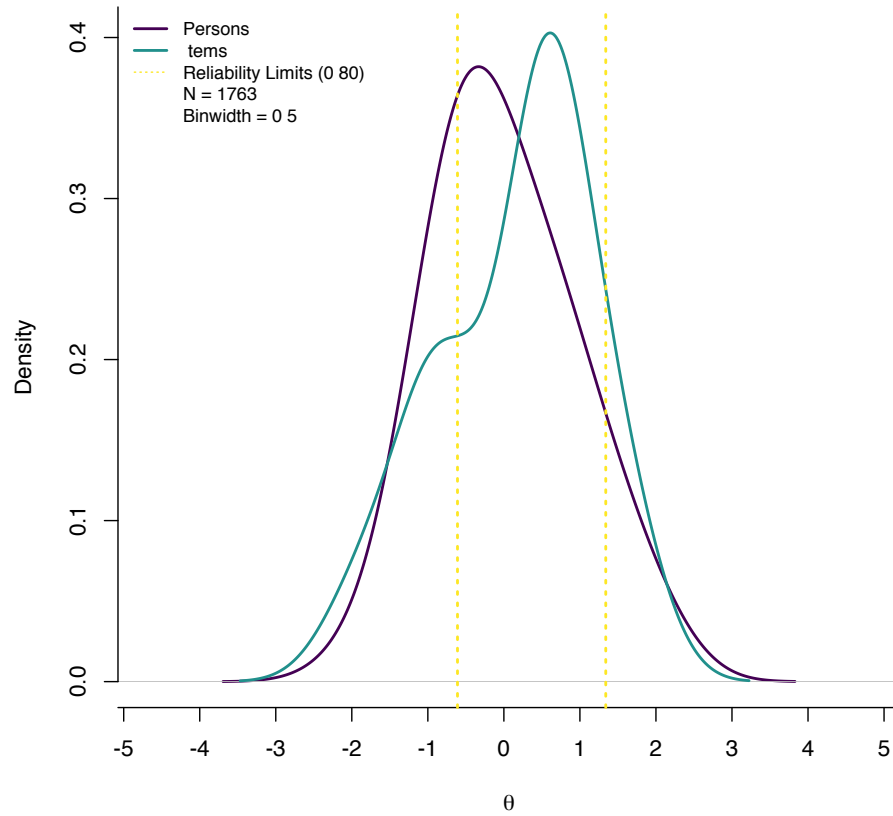


Figure 4.7. Grade 1 item-person plot.

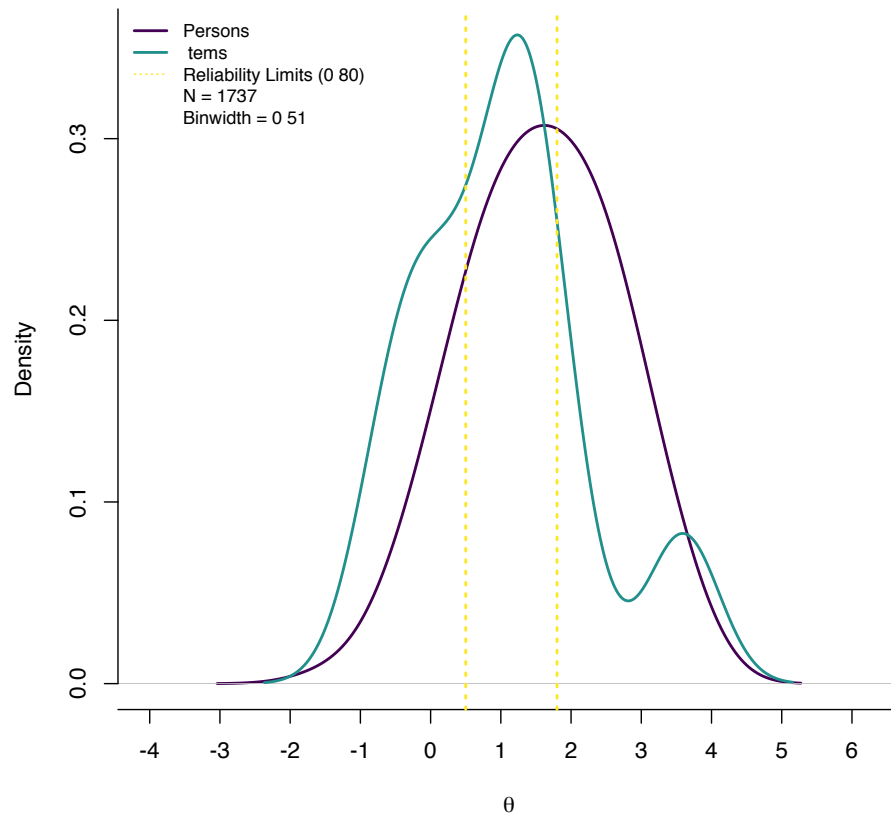


Figure 4.8. Grade 2 item-person plot.

#### 4.4. Predictive Validity

A regression model was used to investigate evidence of predictive validity, with the Fall 2015 K–2 EMSA scores predicting the Spring 2016 K–2 EMSA. We note that the fall and spring scores used in these analyses were not equated. Only students who had scores in both fall and spring were included. Descriptive statistics for the fall and spring samples, split by grade level, are provided in Table 4.7.

*Table 4.7. Sample Descriptives for the Ability Estimates Generated by the Fall 2015 K–2 EMSA and Spring 2016 K–2 EMSA Tests, Split by Grade Level (Students with both Fall and Spring Scores Only)*

Grade level	Number of students	Mean	Standard deviation
<i>Fall 2015 K–2 EMSA</i>			
K	772	–2.037	0.785
1	1,475	0.067	0.930
2	1,400	1.629	1.070
<i>Spring 2016 K–2 EMSA</i>			
K	772	–0.969	0.459
1	1,475	0.041	0.914
2	1,400	0.548	0.976

*Note.* These statistics are limited to students in the sample with both fall and spring scores. The two EMSA tests are different tests; they were vertically equated across grade levels within each season (i.e., fall, spring), but the tests are not equated across seasons, so the fall and spring sample mean ability estimates are not comparable.

On the basis of a sample of 3,647 grade K–2 students who completed both the fall and spring test, and using SPSS version 24, we found a Pearson correlation of .721 ( $p < .001$ ) between the ability estimates generated by the Fall 2015 K–2 EMSA test data and the ability estimates. Therefore, with no adjustment for other factors such as clustering in schools, the student ability estimates from the Fall 2015 K–2 EMSA explains approximately 52% of the variance in student scores measured at the end of the school year for these K–2 students.

Splitting the sample by grade level, 772, 1,475, and 1,400 students represent grades K, 1, and 2, respectively. Again using SPSS version 24, we found a Pearson correlation coefficient of .478 for the grade K sample, .554 for the grade 1 sample, and .630 for the grade 2 sample. All correlations were statistically significant ( $p < .001$ ).

## 5. Discussion and Reflection

The fall 2015 field test was a first attempt at creating the vertically equated tests across the three grade levels. This task faced the challenge of balancing the overall length of the test with the number of anchor items used to link adjacent grade levels. Teachers did not complain about the length of the tests, so the feasibility test results indicate the tests fit into the school program reasonably well. The number of items and the selected-response format seemed to be acceptable for each grade level.

The results of data analysis and feedback from teachers revealed that the difficulty of the grade K test was too high and that it did not include enough moderate-difficulty items. Future versions should include more low- and moderate-difficulty items. Some of the items on the grade K test form were removed during data analysis, leaving space for replacement items on future forms.

The difficulty and reliability of the grade 1 test form appeared to be adequate. Future versions should incorporate more low-difficulty items at grade 1. These items may or may not be useable as anchor items to link grades K and 1.

The grade 2 test form might be too easy for beginning-of-year second-grade students. Several items with high difficulty estimates should be added to the grade 2 test form to improve reliability and overall quality of measurement.

Test reliability appears to be sufficiently high for most of the grade 1 and 2 samples. Grade K reliability might be improved in future versions by replacement items that were removed during the data analysis process with items of moderate difficulty for beginning-of-grade K students.

The fall 2015 field test represented our first attempt in using the optical scanning software for data entry. Use of the software introduced some minor formatting constraints. Staples had to be removed from test packets for scanning, and student responses had to be reviewed page-by-page to be sure they would be identified correctly by the software. Use of software resulted in lower efficiency than did manual data entry. Additionally, the use of the scanner required heavier-weight paper, and the cost of the heavier-weight paper was higher than the lighter-weight paper.

Overall, the content review, feasibility study, and results of data analysis indicate the Fall 2015 EMSA tests provided an adequate assessment tool for its intended purpose.

## 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 (2<sup>nd</sup> Ed.)*. Portsmouth, NH: Heinemann.
- Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.
- Dunbar, S. B., Hoover, H. D., Frisbie, D. A., Ordman, V. L., Oberley, K. R., Naylor, R. J., and Bray, G. B. (2008). *Iowa Test of Basic Skills®*, Form C, Level 7. Rolling Meadows, IL: Riverside Publishing.
- FileMaker Pro (Version 14.1) [Computer Software]. Filemaker, Inc.
- 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>
- Remark Office OMR (Service Pack 4) [Computer Software]. Gravic, Inc.
- 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 Research*, 14, 403–414. [http://dx.doi.org/10.1207/s15327906mbr1404\\_2](http://dx.doi.org/10.1207/s15327906mbr1404_2)
- Schoen, R. C., Anderson, D., & Bauduin, C. (2017). Elementary mathematics student assessment: Measuring the performance of grade K, 1, and 2 students in counting, word problems, and computation in spring 2016. (Research Report No. 2017-22.) Tallahassee, FL: Learning Systems Institute, Florida State University.
- Schoen, R. C., LaVenía, 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., LaVenía, 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., LaVenía, M., Champagne, Z. M., Farina, K., & Tazaz, A. (2016). 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>
- Schoen, R. C., LaVenía, M., Champagne, Z. M., & Farina, K., (2016). Mathematics performance and cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2014. (Research Report No. 2016-01). Tallahassee, FL: Florida Center for Research in Science, Technology, Engineering, and Mathematics. <http://dx.doi.org/doi:10.1725/fsu.1493238156>
- Schoen, R. C., Champagne, Z. M., Whitacre, I., & McCrackin, S. (in review). Comparing the frequency and variation of additive word problems in U.S. first-grade textbooks in the 1980s and the Common Core era. *Teachers College Record*.
- Stigler, J. W., Fuson, K. C., Ham, M., & Kim, M. S. (1986). An analysis of addition and subtraction word problems in American and Soviet elementary mathematics textbooks. *Cognition and Instruction*, 3(3), 153–171.
- 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 fall 2015. As a result, no headers or footers are used in this section of the report.



**Kindergarten – Beginning of Year  
Student Mathematics Assessment**

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

**Sample fill in the bubble multiple-choice:**

What grade are you in?

- |                                  |                       |                       |                       |                       |
|----------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| K                                | 1                     | 2                     | 3                     | 4                     |
| <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

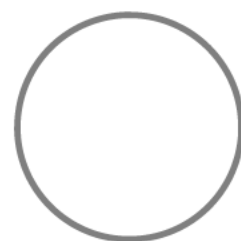


FSU use only - 0815

Affix Barcode Here



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





0



0





0



0



0



0



0





0



0



0

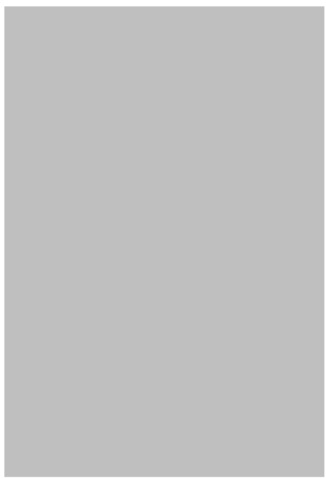


0

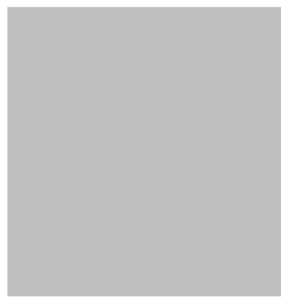


0

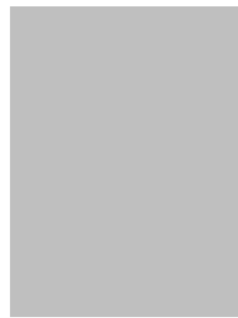




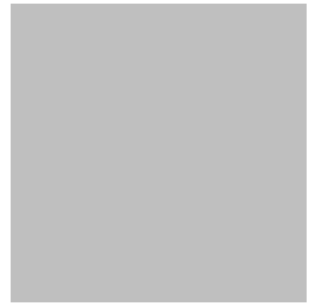
0



0



0



0





0



0



0



0



0





0



0



0



0



0







0



0



0



0



0





0



0



0



0



0





0



0



0

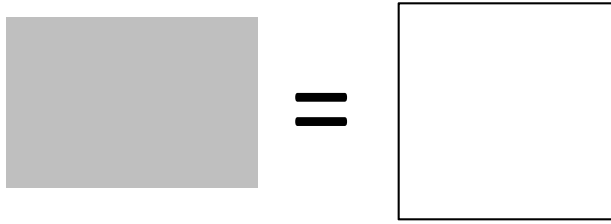


0



0





0



0



0

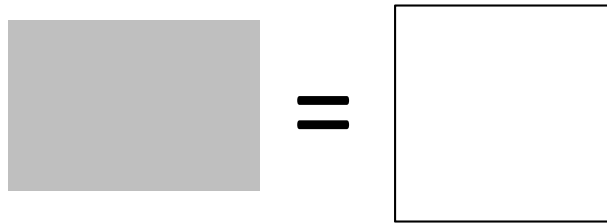


0



0





0



0



0

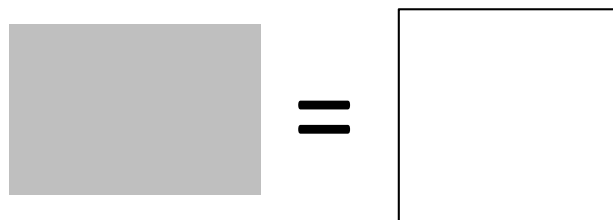
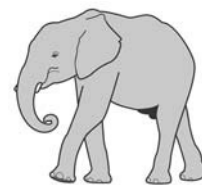


0



0





0



0



0

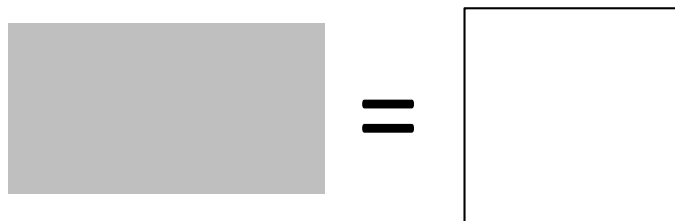


0



0





0



0



0



0



0



[This page was intentionally left blank]

Copyright 2015, Florida State University. The items in this assessment may not be reproduced or used without written consent of Dr. Robert C. Schoen, Associate Director, Florida Center for Research in Science, Technology, Engineering, and Mathematics, Learning Systems Institute, Florida State University (rschoen@lsi.fsu.edu).

*Note.* All used and unused test booklets and administration guides are to be returned to FSU in the same packaging materials in which they arrived. If you have any questions about test administration or materials pick-up, please contact Dr. Amanda Tazaz, atazaz@lsi.fsu.edu.





## Appendix B. Grade 1 Test

The form in this appendix is identical to the form used in fall 2015. As a result, no headers or footers are used in this section of the report.

**First Grade – Beginning of Year  
Student Mathematics Assessment**

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

**Sample fill in the bubble multiple-choice:**

What grade are you in?

K	1	2	3	4
○	●	○	○	○



FSU use only - 0815  
Affix Barcode Here

[This page was intentionally left blank]

Copyright 2015, Florida State University. The items in this assessment may not be reproduced or used without written consent of Dr. Robert C. Schoen, Associate Director, Florida Center for Research in Science, Technology, Engineering, and Mathematics, Learning Systems Institute, Florida State University (rschoen@lsi.fsu.edu).

*Note.* All used and unused test booklets and administration guides are to be returned to FSU in the same packaging materials in which they arrived. If you have any questions about test administration or materials pick-up, please contact Dr. Amanda Tazaz, atazaz@lsi.fsu.edu.





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0







0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0

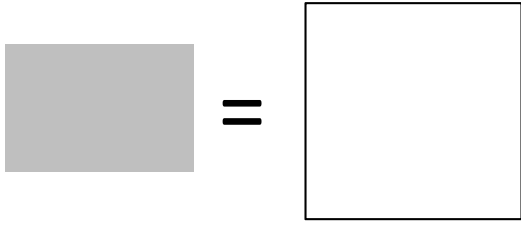


0



0





0



0



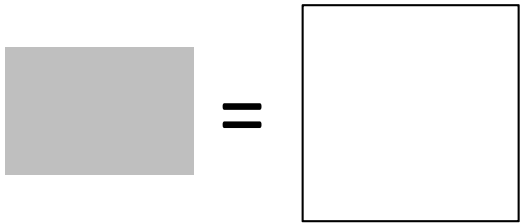
0



0



0



0



0



0

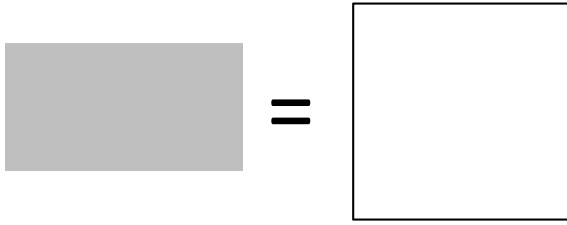


0



0





0



0



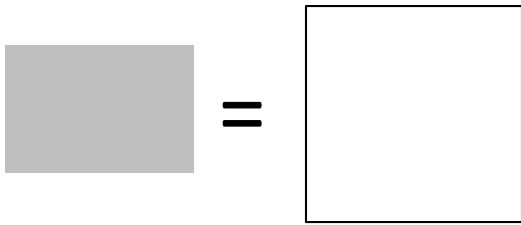
0



0



0



0



0



0



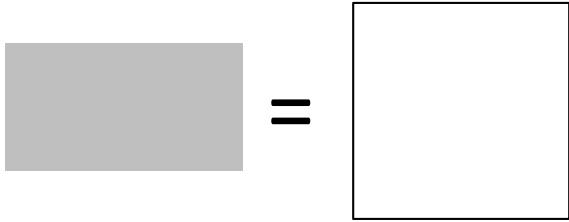
0



0







0



0



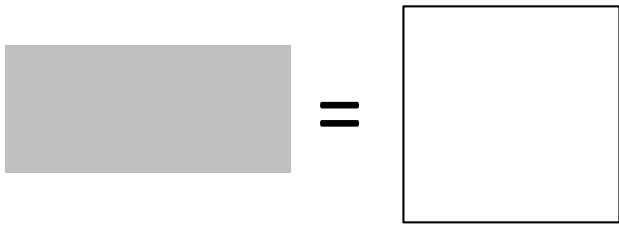
0



0



0



0



0



0

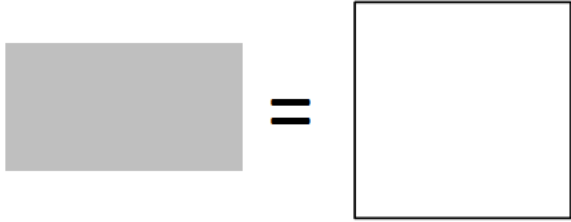


0



0





0



0



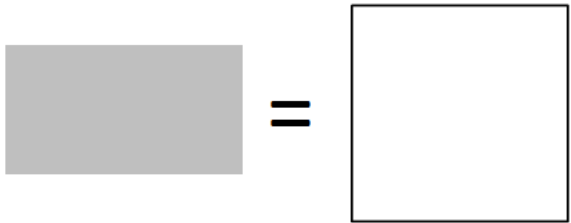
0



0



0



0



0



0

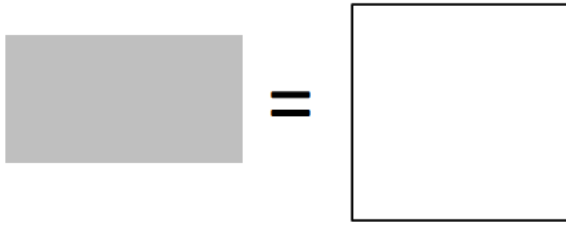


0



0





0



0



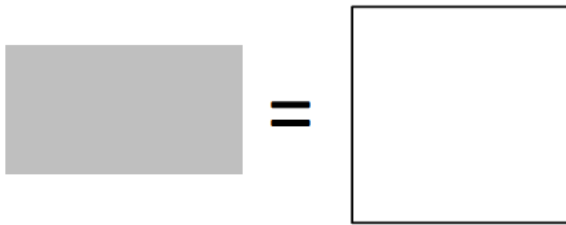
0



0



0



0



0



0



0



0



## Appendix C. Grade 2 Test

The form in this appendix is identical to the form used in fall 2015. As a result, no headers or footers are used in this section of the report.

**Second Grade – Beginning of Year  
Student Mathematics Assessment**

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

**Sample fill in the bubble multiple-choice:**

What grade are you in?

- |                       |                       |                                  |                       |                       |
|-----------------------|-----------------------|----------------------------------|-----------------------|-----------------------|
| K                     | 1                     | 2                                | 3                     | 4                     |
| <input type="radio"/> | <input type="radio"/> | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> |



FSU use only - 0815  
Affix Barcode Here

[This page was intentionally left blank]

Copyright 2015, Florida State University. The items in this assessment may not be reproduced or used without written consent of Dr. Robert C. Schoen, Associate Director, Florida Center for Research in Science, Technology, Engineering, and Mathematics, Learning Systems Institute, Florida State University (rschoen@lsi.fsu.edu).

*Note.* All used and unused test booklets and administration guides are to be returned to FSU in the same packaging materials in which they arrived. If you have any questions about test administration or materials pick-up, please contact Dr. Amanda Tazaz, atazaz@lsi.fsu.edu.





0



0



0



0



0





0



0



0



0



0







0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



0



0





0



0



0



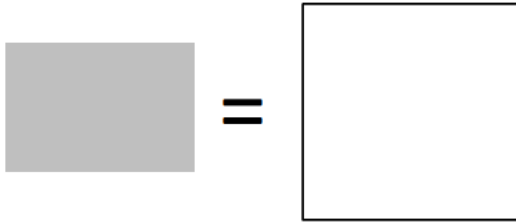
0



0







0



0



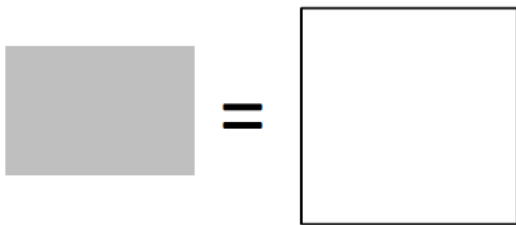
0



0



0



0



0



0

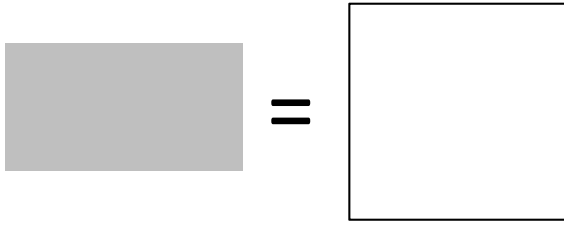


0



0





0



0



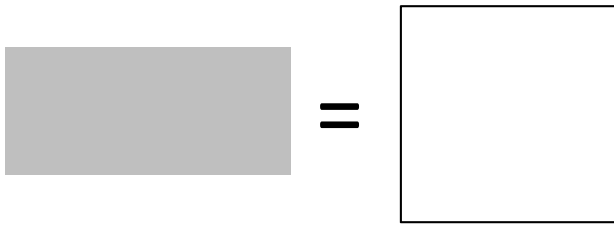
0



0



0



0



0



0

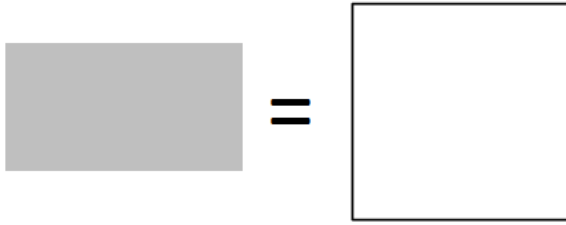


0



0





0



0



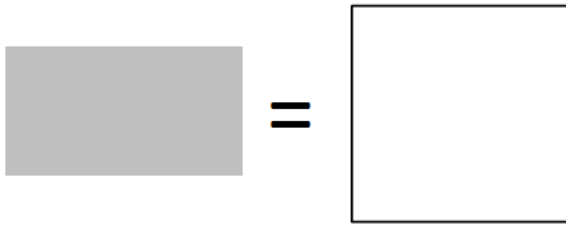
0



0



0



0



0



0

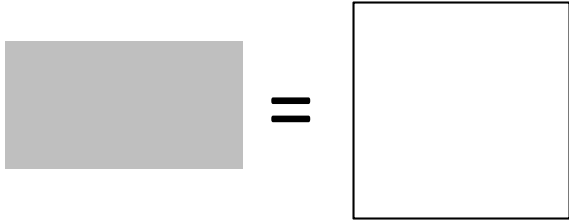


0



0





0



0



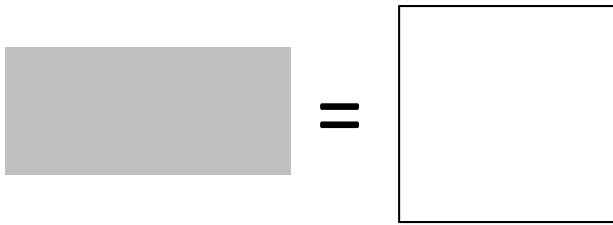
0



0



0



0



0



0

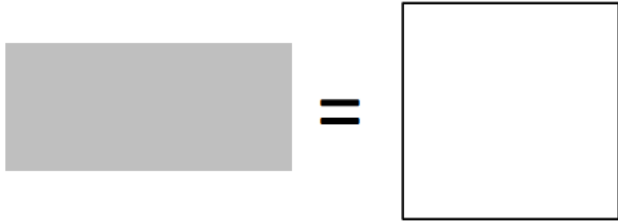


0



0





0



0



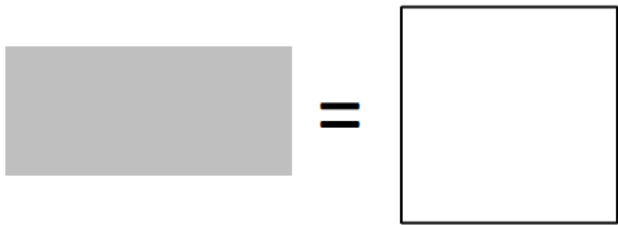
0



0



0



0



0



0



0



0



## Appendix D. Grade K Administration Guide

The form in this appendix is identical to the form used in fall 2015. 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  
Beginning of Year Student Mathematics Assessment**

**August 2015–2016**

**Copyright 2015, Florida State University. Not for reproduction or use without written consent of Dr. Robert C. Schoen, *Foundations for Success in STEM* principal investigator. Instrument development supported by the Florida Department of Education through the U. S. Department of Education Math-Science Partnership program, grant award # 371-2355B-5C001.**

## 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 Beginning of the Year Student Assessment during the first two weeks of school. If you cannot administer the assessment during that window, please notify Amanda Tazaz ([atazaz@lsi.fsu.edu](mailto:atazaz@lsi.fsu.edu)) and plan to administer it as early as possible in the school year.

You will notice that the assessment contains three basic sections: Counting, Word Problems, and Computation. All items on the test use a multiple-choice format. We ask that students use pencils to bubble their answers. A requested 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 assessment. At the end of this document, we have enclosed a blank roster form 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 Beginning of the Year Student Assessment Materials section on page 4.

## Beginning of the Year Student Assessment Window

Student testing will occur according to the following schedule:

<u>School District</u>	<u>Testing Window</u>
District 1	August 18, 2015 – August 28, 2015
District 2	August 24, 2015 – September 4, 2015
District 3	August 17, 2015 – August 28, 2015
District 4	August 12, 2015 – August 26, 2015
District 5	August 10, 2015 – August 24, 2015
District 6	August 17, 2015 – August 28, 2015
District 7	August 17, 2015 – August 28, 2015
District 8	August 24, 2015 – September 4, 2015
District 9	August 17, 2015 – August 28, 2015
District 10	August 10, 2015 – August 24, 2015
District 11	August 10, 2015 – August 24, 2015
District 12	August 20, 2015 – September 3, 2015
District 13	August 10, 2015 – August 24, 2015

## Materials

The following materials are required for testing:

- Beginning of the Year Student Assessment Guidelines and Administration Instructions (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](mailto: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 assessment to students for whom you are the teacher of record. Therefore, if you teach multiple groups of students mathematics, you only need to administer the assessment with students that 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:

Prior to the testing session, the classroom teacher must enter this information (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 it for students to enter this information.)

The Beginning of the Year Student Assessment for the *Foundations for Success in STEM* Study may be administered to students on either an individual or 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 pre-labeled test booklets are properly paired (i.e., each student receives the test booklet that has his or her name written on it).
- Provide students with a comfortable testing environment.
- Testing administrators should adhere to the Beginning of the Year Student Assessment guidelines and administration instructions.
- No talking or communication between students is permitted during testing.
- The test is intended to be read aloud to students by the testing administrator.
- Students are permitted to use mathematics manipulatives during the pre-test if they would ordinarily be permitted to use manipulatives in your classroom.
- The administration script indicates that teachers should read the question 2 times. However, it is permissible for teachers to read the problem more than 2 times if needed.

### **Administering the Beginning of the Year Student Assessment**

It is assumed that the classroom teacher will administer the assessment; however, other school personnel (such as a paraprofessional or even a substitute teacher) can administer the assessment, providing they follow the assessment protocol as described below.

The testing conditions for the Beginning of the Year Student 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.

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, it is important that appropriate procedures are followed when administering the Beginning of the Year Student Assessment. These procedures include:

- Administration of the appropriate test level (Kindergarten assessment for Grade K students, etc.)
- Adherence to the Beginning of the Year Student 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 pre-test should take approximately 45 minutes. This is not a timed test, and students should be allowed adequate time to answer the test questions.

### **Submitting the Beginning of the Year Student 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 pre-test guidelines and administration instruction document and your completed student information sheet in the package. A member of the project will coordinate with your school to set a date to retrieve the testing materials from you.

The target period of pickup of material will be as follows (you will receive an email prior to pick-up to ensure the material is ready in the front office).

<b><u>School District</u></b>	<b><u>Target Pick-up Window</u></b>
District 1	August 31, 2015 – September 4, 2015
District 2	September 7, 2015 – September 11, 2015
District 3	August 31, 2015 – September 4, 2015
District 4	August 27, 2015 – September 3, 2015
District 5	August 24, 2015 – August 28, 2015
District 6	August 31, 2015 – September 4, 2015
District 7	August 31, 2015 – September 4, 2015
District 8	September 7, 2015 – September 11, 2015
District 9	August 31, 2015 – September 4, 2015
District 10	August 24, 2015 – August 28, 2015
District 11	August 24, 2015 – August 28, 2015
District 12	September 7, 2015 – September 11, 2015
District 13	August 24, 2015 – August 28, 2015

If you have questions about this process, contact [atazaz@lsi.fsu.edu](mailto:atazaz@lsi.fsu.edu) .

## Pre-test Administration Instructions – Kindergarten

[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 all students have a pencil.

I will now pass out the assessments. The assessments are already labeled with your names. When you receive the assessment, keep it face up, and do not turn any pages; we will all begin at the same time after I go over the instructions. 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., each student receives the test booklet that has his or her name written on it).

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

The 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 underneath 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 all students have filled in the bubble under the triangle.

For each question, 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 can use the white space on the paper to work out your answers. Please do not mark on the barcode at the bottom of each page.

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.

Again:

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.

? Fill in the bubble under the correct answer.

Again: ? Fill in the bubble under the correct answer.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the smiley face at the top.

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

[REDACTED] ? Fill in the bubble under the correct answer.

Again: [REDACTED] ? Fill in the bubble under the correct answer.

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.

[REDACTED]

Again: [REDACTED].

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.

[REDACTED] ? Fill in the bubble under the correct answer.

Again: [REDACTED] ? Fill in the bubble under the correct answer.

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.

[REDACTED] ? Fill in the bubble under the correct answer.

Again: [REDACTED] ? Fill in the bubble under the correct answer.

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.

[REDACTED]

Fill in the bubble under the correct answer.

Again: [REDACTED]  
[REDACTED]


Fill in the bubble under the correct answer.

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.



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

Again:




Fill in the bubble under 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.



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

Again:



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

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the fish at the top.

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

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

Again:

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

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the zebra at the top.

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

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

Again:

Fill in the bubble under 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 elephant at the top.

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



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

Again:

Fill in the bubble under 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 house at the top.

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

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

Again:

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

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

You have completed the assessment. Thank you for working hard and trying your best. Please close your test book, and I will collect it.

Collect all testing materials.





## Appendix E. Grade 1 Administration Guide

The form in this appendix is identical to the form used in fall 2015. 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  
Beginning of Year Student Mathematics Assessment**

**August 2015–2016**

**Copyright 2015, Florida State University. Not for reproduction or use without written consent of Dr. Robert C. Schoen, *Foundations for Success in STEM* principal investigator. Instrument development supported by the Florida Department of Education through the U. S. Department of Education Math-Science Partnership program, grant award # 371-2355B-5C001.**

## 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 pretesting 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 Beginning of the Year Student Assessment during the first two weeks of school. If you cannot administer the assessment during that window, please notify Amanda Tazaz ([atazaz@lsi.fsu.edu](mailto:atazaz@lsi.fsu.edu)) and administer the assessment as early as possible in the school year.

You will notice that the assessment contains three basic sections: Counting, Word Problems, and Computation. All items on the test use a multiple-choice format. Students may use markers or pencils to bubble their answers. A requested 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 assessment. At the end of this document, we have enclosed a blank roster form 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 Beginning of the Year Student Assessment Materials section on page 4.

## Beginning of the Year Student Assessment Window

Student testing will occur according to the following schedule:

<b>School District</b>	<b>Testing Window</b>
District 1	August 18, 2015 – August 28, 2015
District 2	August 24, 2015 – September 4, 2015
District 3	August 17, 2015 – August 28, 2015
District 4	August 12, 2015 – August 26, 2015
District 5	August 10, 2015 – August 24, 2015
District 6	August 17, 2015 – August 28, 2015
District 7	August 17, 2015 – August 28, 2015
District 8	August 24, 2015 – September 4, 2015
District 9	August 17, 2015 – August 28, 2015
District 10	August 10, 2015 – August 24, 2015
District 11	August 10, 2015 – August 24, 2015
District 12	August 20, 2015 – September 3, 2015
District 13	August 10, 2015 – August 24, 2015

## Materials

The following materials are required for testing:

- Beginning of the Year Student Assessment Guidelines and Administration Instructions (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](mailto: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 assessment to students for whom you are the teacher of record. Therefore, if you teach multiple groups of students mathematics, you only need to administer the assessment with students that 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:

Prior to the testing session, the classroom teacher must enter this information (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 it for students to enter this information.)

The Beginning of the Year Student Assessment for the *Foundations for Success in STEM* Study may be administered to students on either an individual or 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 pre-labeled test booklets are properly paired (i.e., each student receives the test booklet that has his or her name written on it).
- Provide students with a comfortable testing environment.
- Testing administrators should adhere to the Beginning of the Year Student Assessment guidelines and administration instructions.
- No talking or communication between students is permitted during testing.
- The test is intended to be read aloud to students by the testing administrator.
- Students are permitted to use mathematics manipulatives during the pre-test if they would ordinarily be permitted to use manipulatives in your classroom.
- The administration script indicates that teachers should read the question 2 times. However, it is permissible for teachers to read the problem more than 2 times if needed.

### **Administering the Beginning of the Year Student Assessment**

It is assumed that the classroom teacher will administer the assessment; however, other school personnel (such as a paraprofessional or even a substitute teacher) can administer the assessment, providing they follow the assessment protocol as described below.

The testing conditions for the Beginning of the Year Student 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.

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, it is important that appropriate procedures are followed when administering the Beginning of the Year Student Assessment. These procedures include:

- Administration of the appropriate test level (Grade 1 assessment for Grade 1 students, etc.).
- Adherence to the Beginning of the Year Student 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 pre-test should take approximately 45 minutes. This is not a timed test, and students should be allowed adequate time to answer the test questions.

### **Submitting the Beginning of the Year Student 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 pre-test guidelines and administration instruction document and your completed student information sheet in the package. A member of the project will coordinate with your school to set a date to retrieve the testing materials from you.

The target period of pickup of material will be as follows (you will receive an email prior to pick-up to ensure the material is ready in the front office).

<b>School District</b>	<b>Target Pick-up Window</b>
District 1	August 31, 2015 – September 4, 2015
District 2	September 7, 2015 – September 11, 2015
District 3	August 31, 2015 – September 4, 2015
District 4	August 27, 2015 – September 3, 2015
District 5	August 24, 2015 – August 28, 2015
District 6	August 31, 2015 – September 4, 2015
District 7	August 31, 2015 – September 4, 2015
District 8	September 7, 2015 – September 11, 2015
District 9	August 31, 2015 – September 4, 2015
District 10	August 24, 2015 – August 28, 2015
District 11	August 24, 2015 – August 28, 2015
District 12	September 7, 2015 – September 11, 2015
District 13	August 24, 2015 – August 28, 2015

If you have questions about this process, contact [atazaz@lsi.fsu.edu](mailto:atazaz@lsi.fsu.edu).



## **Pre-test Administration Instructions – First 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 all students have a pencil.

I will now pass out the assessments. The assessments are already labeled with your names. When you receive the assessment, keep it face up, and do not turn any pages; we will all begin at the same time after I go over the instructions. 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., each student receives the test booklet that has his or her name written on it).

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

The 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 1, for first grade. Notice how the bubble beneath (below) the 1 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.

For each question, 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 can use the white space on the paper to work out your answers. Please do not mark on the barcode at the bottom of each page.

Are there any questions?

Address any questions.

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

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

 ? Fill in the bubble under the correct answer.

Again:  ? Fill in the bubble under the correct answer.

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 under the correct answer.

Again:  ? Fill in the bubble under the correct answer.

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 under the correct answer.

Again: \_\_\_\_\_ ? Fill in the bubble under the correct answer.

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 under the correct answer.

Again: \_\_\_\_\_ ? Fill in the bubble under the correct answer.

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.

\_\_\_\_\_ ? Fill in the bubble under the correct answer.

Again: \_\_\_\_\_ ? Fill in the bubble under the correct answer.

When you finish, 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

[Redacted]

[Redacted] ?

Fill in the bubble under the correct answer.

Again: [Redacted]

[Redacted]

Fill in the bubble under the correct answer.

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.

[Redacted]

Fill in the bubble under the correct answer.

Again: [Redacted]

[Redacted]

Fill in the bubble under the correct answer.

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.

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

Again:

Fill in the bubble under 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 under the correct answer.

Again:

Fill in the bubble under the correct answer.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the zebra at the top.

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

[REDACTED]

Fill in the bubble under the correct answer.

Again: [REDACTED]

Fill in the bubble under the correct answer.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

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

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

[REDACTED]

Fill in the bubble under the correct answer.

Again: [REDACTED]

Fill in the bubble under the correct answer.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the fish 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 five pages have some addition and subtraction problems that you will solve at your own pace.

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 fall 2015. 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  
Beginning of Year Student Mathematics Assessment**

**August 2015–2016**

**Copyright 2015, Florida State University. Not for reproduction or use without written consent of Dr. Robert C. Schoen, *Foundations for Success in STEM* principal investigator. Instrument development supported by the Florida Department of Education through the U. S. Department of Education Math-Science Partnership program, grant award # 371-2355B-5C001.**

## 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 pretesting 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 Beginning of the Year Student during the first two weeks of school. If you cannot administer the assessment during that window, please notify Amanda Tazaz ([atazaz@lsi.fsu.edu](mailto:atazaz@lsi.fsu.edu)) and administer the assessment as early as possible in the school year.

You will notice that the assessment contains three basic sections: Counting, Word Problems, and Computation. All items on the test use a multiple-choice format. Students may use markers or pencils to bubble their answers. A requested 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 assessment. At the end of this document, we have enclosed a blank roster form 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 Beginning of the Year Student Assessment Materials section on page 4.

## Beginning of the Year Student Assessment Window

Student testing will occur according to the following schedule:

<b>School District</b>	<b>Testing Window</b>
District 1	August 18, 2015 – August 28, 2015
District 2	August 24, 2015 – September 4, 2015
District 3	August 17, 2015 – August 28, 2015
District 4	August 12, 2015 – August 26, 2015
District 5	August 10, 2015 – August 24, 2015
District 6	August 17, 2015 – August 28, 2015
District 7	August 17, 2015 – August 28, 2015
District 8	August 24, 2015 – September 4, 2015
District 9	August 17, 2015 – August 28, 2015
District 10	August 10, 2015 – August 24, 2015
District 11	August 10, 2015 – August 24, 2015
District 12	August 20, 2015 – September 3, 2015
District 13	August 10, 2015 – August 24, 2015

## Materials

The following materials are required for testing:

- Beginning of the Year Student Assessment Guidelines and Administration Instructions (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](mailto: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 assessment to students for whom you are the teacher of record. Therefore, if you teach multiple groups of students mathematics, you only need to administer the assessment with students that 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:

Prior to the testing session, the classroom teacher must enter this information (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 it for students to enter this information.)

The Beginning of the Year Student Assessment for the *Foundations for Success in STEM* Study may be administered to students on either an individual or 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 pre-labeled test booklets are properly paired (i.e., each student receives the test booklet that has his or her name written on it).
- Provide students with a comfortable testing environment.
- Testing administrators should adhere to the Beginning of the Year Student Assessment guidelines and administration instructions.
- No talking or communication between students is permitted during testing.
- The test is intended to be read aloud to students by the testing administrator.
- Students are permitted to use mathematics manipulatives during the pre-test if they would ordinarily be permitted to use manipulatives in your classroom.
- The administration script indicates that teachers should read the question 2 times. However, it is permissible for teachers to read the problem more than 2 times if needed.

### **Administering the Beginning of the Year Student Assessment**

It is assumed that the classroom teacher will administer the assessment; however, other school personnel (such as a paraprofessional or even a substitute teacher) can administer the assessment, providing they follow the assessment protocol as described below.

The testing conditions for the Beginning of the Year Student 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.

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, it is important that appropriate procedures are followed when administering the Beginning of the Year Student Assessment. These procedures include:

- Administration of the appropriate test level (Second Grade assessment for Second Grade students, etc.)
- Adherence to the Beginning of the Year Student 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 pre-test should take approximately 45 minutes. This is not a timed test, and students should be allowed adequate time to answer the test questions.

### **Submitting the Beginning of the Year Student 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 pre-test guidelines and administration instruction document and your completed student information sheet in the package. A member of the project will coordinate with your school to set a date to retrieve the testing materials from you.

The target period of pickup of material will be as follows (you will receive an email prior to pick-up to ensure the material is ready in the front office):

<b>School District</b>	<b>Target Pick-up Window</b>
District 1	August 31, 2015 – September 4, 2015
District 2	September 7, 2015 – September 11, 2015
District 3	August 31, 2015 – September 4, 2015
District 4	August 27, 2015 – September 3, 2015
District 5	August 24, 2015 – August 28, 2015
District 6	August 31, 2015 – September 4, 2015
District 7	August 31, 2015 – September 4, 2015
District 8	September 7, 2015 – September 11, 2015
District 9	August 31, 2015 – September 4, 2015
District 10	August 24, 2015 – August 28, 2015
District 11	August 24, 2015 – August 28, 2015
District 12	September 7, 2015 – September 11, 2015
District 13	August 24, 2015 – August 28, 2015

If you have questions about this process, contact [atazaz@lsi.fsu.edu](mailto:atazaz@lsi.fsu.edu).

## **Pre-test 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 all students have a pencil.

I will now pass out the assessments. The assessments are already labeled with your names. When you receive the assessment, keep it face up, and do not turn any pages; we will all begin at the same time after I go over the instructions. 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., each student receives the test booklet that has his or her name written on it).

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

The 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 2, for second grade. Notice how the bubble beneath (below) the 2 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.

For each question, 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 can use the white space on the paper to work out your answers. Please do not mark on the barcode at the bottom of each page.

Are there any questions?

Address any questions.

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 under the correct answer.

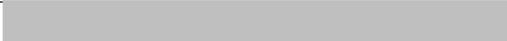
Again:  ? Fill in the bubble under the correct answer.

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 under the correct answer.

Again:  ? Fill in the bubble under the correct answer.

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.

[REDACTED] ? Fill in the bubble under the correct answer.

Again: [REDACTED] ? Fill in the bubble under the correct answer.

Pause and wait for all students to complete the item.

Turn to the page with the zebra at the top.

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

[REDACTED]  
Fill in the bubble under the correct answer.

Again [REDACTED]  
[REDACTED]


Fill in the bubble under the correct answer.

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.



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

Again:




Fill in the bubble under 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 under the correct answer.

Again:



Fill in the bubble under the correct answer.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

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

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

[REDACTED]

Fill in the bubble under the correct answer.

Again:

[REDACTED]

Fill in the bubble under the correct answer.

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.

[REDACTED]

Fill in the bubble under the correct answer.

Again:

[REDACTED]


Fill in the bubble under the correct answer.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.


Turn to the page with the smiley face at the top.

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



Fill in the bubble under the correct answer.

Again:




Fill in the bubble under the correct answer.

Pause and wait for all students to complete the item.


Turn to the page with the balloons at the top.

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



Fill in the bubble under the correct answer.

Again:



Fill in the bubble under the correct answer.

When you are finished, put your pencil down.

Pause and wait for all students to complete the item.

Turn to the page with the fish 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 five pages have some addition and subtraction problems that you will solve at your own pace.

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. Grade K Scoring Key*





























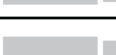



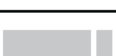





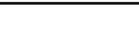

Item	Item description	Data entry	Correct response
GKi2		Record number corresponding with student's response, DNS, or UI	
GKi3		Record number corresponding with student's response, DNS, or UI	
GKi4_G1i1		Record number corresponding with student's response, DNS, or UI	
GKi5		Record number corresponding with student's response, DNS, or UI	
GKi6_G1i2		Record number corresponding with student's response, DNS, or UI	
GKi7_G1i3_G2i1		Record number corresponding with student's response, DNS, or UI	
GKi8_G1i6		Record number corresponding with student's response, DNS, or UI	
GKi9_G1i7		Record number corresponding with student's response, DNS, or UI	
GKi10_G1i8_G2i5		Record number corresponding with student's response, DNS, or UI	
GKG2i11_G1i12		Record number corresponding with student's response, DNS, or UI	
GKG2i12_G1i13		Record number corresponding with student's response, DNS, or UI	
GKi13_G1i15		Record number corresponding with student's response, DNS, or UI	
GKi14_G1i16_G2i13		Record number corresponding with student's response, DNS, or UI	



Table G.2. Grade 1 Scoring Key

Item	Item description	Data entry	Correct response
GKi4_G1i1		Record number corresponding with student's response, DNS, or UI	
GKi6_G1i2		Record number corresponding with student's response, DNS, or UI	
GKi7_G1i3_G2i1		Record number corresponding with student's response, DNS, or UI	
G1i4_G2i2		Record number corresponding with student's response, DNS, or UI	
G1i5_G2i3		Record number corresponding with student's response, DNS, or UI	
GKi8_G1i6		Record number corresponding with student's response, DNS, or UI	
GKi9_G1i7		Record number corresponding with student's response, DNS, or UI	
GKi10_G1i8_G2i5		Record number corresponding with student's response, DNS, or UI	
G1i9_G2i6		Record number corresponding with student's response, DNS, or UI	
G1i10_G2i4		Record number corresponding with student's response, DNS, or UI	
G1i11_G2i7		Record number corresponding with student's response, DNS, or UI	
GKG2i11_G1i12		Record number corresponding with student's response, DNS, or UI	
GKG2i12_G1i13		Record number corresponding with student's response, DNS, or UI	
G1i14_G2i16		Record number corresponding with student's response, DNS, or UI	
GKi13_G1i15		Record number corresponding with student's response, DNS, or UI	
GKi14_G1i16_G2i13		Record number corresponding with student's response, DNS, or UI	
G1i17_G2i14		Record number corresponding with student's response, DNS, or UI	
G1i18		Record number corresponding with student's response, DNS, or UI	
G1i19		Record number corresponding with student's response, DNS, or UI	
G1i20		Record number corresponding with student's response, DNS, or UI	
G1i21_G2i17		Record number corresponding with student's response, DNS, or UI	

Table G.3. Grade 2 Scoring Key

Item	Item description	Data entry	Correct response
GKi7_G1i3_G2i1		Record number corresponding with student's response, DNS, or UI	
G1i4_G2i2		Record number corresponding with student's response, DNS, or UI	
G1i5_G2i3		Record number corresponding with student's response, DNS, or UI	
G1i10_G2i4		Record number corresponding with student's response, DNS, or UI	
GKi10_G1i8_G2i5		Record number corresponding with student's response, DNS, or UI	
G1i9_G2i6		Record number corresponding with student's response, DNS, or UI	
G1i11_G2i7		Record number corresponding with student's response, DNS, or UI	
G2i8		Record number corresponding with student's response, DNS, or UI	
G2i9		Record number corresponding with student's response, DNS, or UI	
G2i10		Record number corresponding with student's response, DNS, or UI	
GKG2i11_G1i12		Record number corresponding with student's response, DNS, or UI	
GKG2i12_G1i13		Record number corresponding with student's response, DNS, or UI	
GKi14_G1i16_G2i13		Record number corresponding with student's response, DNS, or UI	
G1i17_G2i14		Record number corresponding with student's response, DNS, or UI	
G2i15		Record number corresponding with student's response, DNS, or UI	
G1i14_G2i16		Record number corresponding with student's response, DNS, or UI	
G1i21_G2i17		Record number corresponding with student's response, DNS, or UI	
G2i18		Record number corresponding with student's response, DNS, or UI	
G2i19		Record number corresponding with student's response, DNS, or UI	
G2i20		Record number corresponding with student's response, DNS, or UI	

## 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. The difficulty and discrimination estimates are based on the vertically scaled models.

Table H.1. Item Statistics for the Grade K Test Based on the Grade K Sample ( $n = 986$ )

Item	Item description	CTT-based statistics		IRT-based statistics	
		PC (se)	PB	Diff (se)	Discrim (se)
GKi2		.90 (.010)	.39	-3.84 (.167)	1.63 (.202)
GKi3		.85 (.011)	.48	-3.23 (.085)	2.64 (.321)
GKi4_G1i1		.73 (.014)	.53	-2.82 (.067)	2.25 (.233)
GKi5		.75 (.014)	.52	-2.99 (.086)	1.68 (.175)
GKi6_G1i2		.36 (.015)	.60	-1.60 (.065)	1.88 (.185)
GKi7_G1i3_G2i1		.28 (.014)	.44	-.82 (.169)	0.87 (.106)
GKi8_G1i6		.39 (.016)	.50	-1.56 (.093)	1.04 (.109)
GKi9_G1i7		.28 (.014)	.47	-.88 (.158)	0.89 (.105)
GKi10_G1i8_G2i5		.09 (.009)	.35	.32 (.303)	1.14 (.156)
GKG2i11_G1i12		.25 (.014)	.43	-.44 (.238)	0.74 (.101)
GKG2i12_G1i13		.18 (.012)	.47	-.52 (.163)	1.24 (.140)
<i>GKi13_G1i15</i>		.21 (.013)	.23	–	–
<i>GKi14_G1i16_G2i13</i>		.14 (.011)	.27	–	–

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 ( $n = 1,763$ )

Item	Item description	CTT-based statistics		IRT-based statistics	
		PC (se)	PB	Diff (se)	Discrim (se)
<i>GKi4_G1i1</i>		.96 (.005)	.18	–	–
GKi6_G1i2		.80 (.009)	.43	-1.46 (.100)	1.22 (.103)
GKi7_G1i3_G2i1		.69 (.011)	.50	-.86 (.069)	1.17 (.088)
G1i4_G2i2		.59 (.012)	.36	-.65 (.105)	.59 (.061)
G1i5_G2i3		.24 (.010)	.38	1.73 (.156)	.75 (.070)
GKi8_G1i6		.78 (.010)	.35	-1.97 (.196)	.72 (.078)
GKi9_G1i7		.72 (.011)	.49	-.97 (.069)	1.27 (.095)
GKi10_G1i8_G2i5		.35 (.011)	.48	.79 (.078)	.93 (.071)
G1i9_G2i6		.25 (.010)	.49	1.24 (.091)	1.05 (.079)
G1i10_G2i4		.39 (.012)	.46	.59 (.076)	.84 (.067)
G1i11_G2i7		.23 (.010)	.45	1.53 (.117)	.93 (.076)
GKG2i11_G1i12		.71 (.011)	.50	-.96 (.073)	1.17 (.091)
GKG2i12_G1i13		.56 (.012)	.47	-.30 (.064)	.90 (.071)
G1i14_G2i16		.36 (.011)	.45	.75 (.081)	.87 (.068)
GKi13_G1i15		.40 (.012)	.60	.30 (.040)	2.01 (.122)
GKi14_G1i16_G2i13		.40 (.012)	.65	.26 (.035)	2.88 (.191)
G1i17_G2i14		.34 (.011)	.65	.48 (.038)	2.55 (.162)
G1i18		.34 (.011)	.63	.48 (.039)	2.42 (.151)
G1i19		.37 (.011)	.63	.39 (.039)	2.36 (.146)
G1i20		.29 (.011)	.57	.77 (.050)	1.80 (.113)
G1i21_G2i17		.29 (.011)	.49	.94 (.069)	1.23 (.084)

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 ( $n = 1,737$ )

Item	Item description	CTT-based statistics		IRT-based statistics	
		PC (se)	PB	Diff (se)	Discrim (se)
GKi7_G1i3_G2i1		.89 (.008)	.43	-.55 (.102)	1.37 (.137)
G1i4_G2i2		.83 (.009)	.46	-.24 (.095)	1.12 (.233)
G1i5_G2i3		.57 (.012)	.56	1.25 (.047)	1.23 (.094)
G1i10_G2i4		.65 (.011)	.56	.92 (.049)	1.36 (.103)
GKi10_G1i8_G2i5		.57 (.012)	.58	1.28 (.044)	1.39 (.104)
G1i9_G2i6		.62 (.012)	.63	1.11 (.040)	1.80 (.134)
G1i11_G2i7		.56 (.012)	.65	1.34 (.038)	1.93 (.145)
G2i8		.52 (.012)	.59	1.49 (.042)	1.43 (.106)
G2i9		.47 (.012)	.60	1.67 (.041)	1.52 (.113)
G2i10		.40 (.012)	.53	2.01 (.048)	1.24 (.096)
<i>GKG2i11_G1i12</i>		<i>.93 (.006)</i>	<i>.37</i>	–	–
GKG2i12_G1i13		.86 (.008)	.40	-.84 (.150)	.90 (.098)
GKi14_G1i16_G2i13		.82 (.009)	.46	-.30 (.106)	.99 (.096)
G1i17_G2i14		.77 (.010)	.54	.26 (.067)	1.25 (.106)
G2i15		.75 (.010)	.46	.00 (.100)	.85 (.082)
G1i14_G2i16		.70 (.011)	.48	.41 (.081)	.86 (.078)
G1i21_G2i17		.63 (.012)	.46	.75 (.077)	.75 (.071)
G2i18		.51 (.012)	.47	1.51 (.061)	.80 (.071)
G2i19		.30 (.011)	.31	3.63 (.220)	.43 (.063)
G2i20		.22 (.010)	.37	3.57 (.144)	.71 (.079)

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 are spaghetti plots based on all of the items on the grade K test by means of a CTT-based approach with some smoothing. Overall, the shapes of the trace lines appear satisfactory.

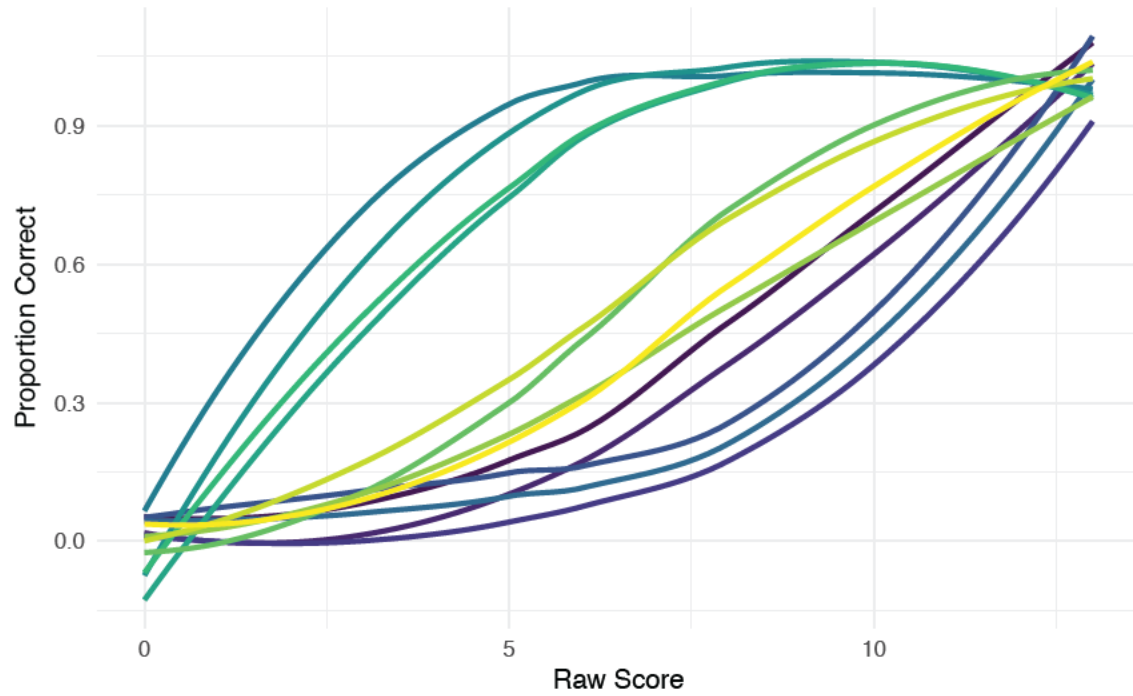


Figure H.1. Grade K spaghetti plot.

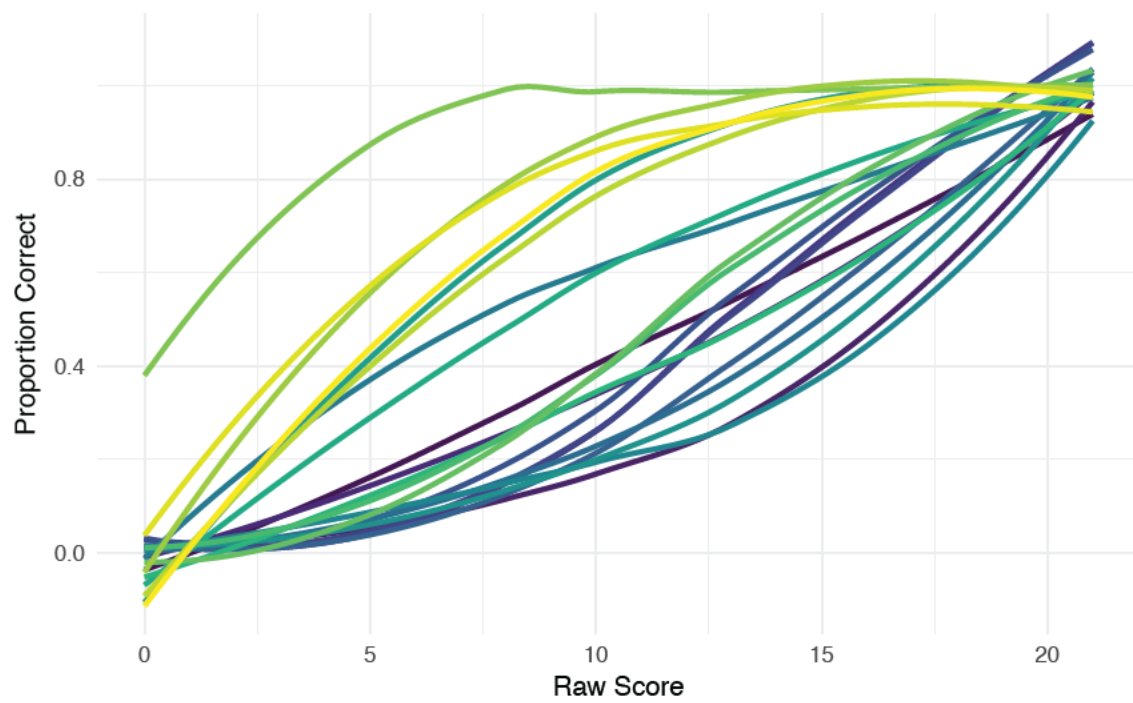


Figure H.2. Grade 1 spaghetti plot.

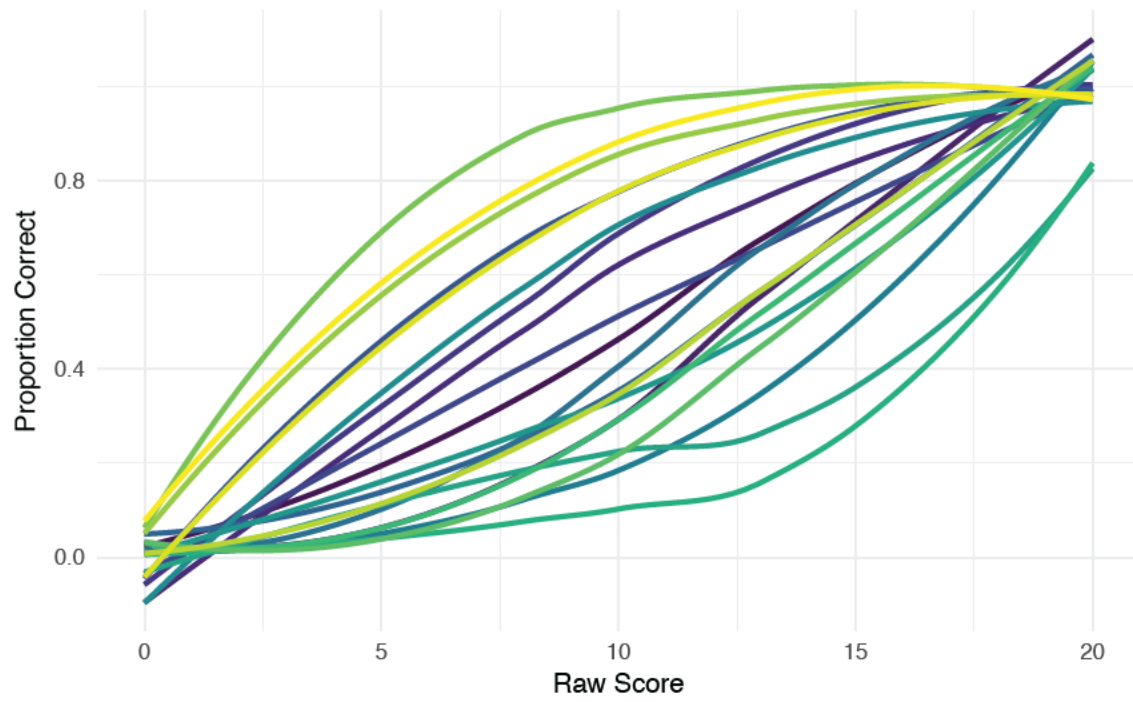


Figure H.3. Grade 2 spaghetti plot.

## 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	Most frequent incorrect responses			
		Response (%)	Response (%)	Response (%)	Response (%)	Response (%)
GKi2		4 (.90)	3 (.05)	DNS (.03)	UI (.02)	–
GKi3		3 (.85)	1 (.05)	5 (.03)	DNS (.02)	UI (.02)
GKi4_G1i1		7 (.73)	6 (.08)	10 (.07)	8 (.05)	DNS (.03)
GKi5		4 (.75)	6 (.13)	3 (.05)	DNS (.03)	1 (.03)
GKi6_G1i2		9 (.36)	10 (.32)	1 (.13)	8 (.09)	7 (.06)
GKi7_G1i3_G2i1		13 (.28)	15 (.43)	1 (.10)	0 (.09)	2 (.07)
GKi8_G1i6		7 (.39)	6 (.16)	4 (.16)	3 (.13)	1 (.10)
GKi9_G1i7		3 (.28)	2 (.23)	4 (.16)	5 (.15)	9 (.13)
GKi10_G1i8_G2i5		8 (.09)	2 (.35)	4 (.24)	12 (.14)	6 (.13)
GKG2i11_G1i12		11 (.25)	7 (.20)	10 (.17)	6 (.17)	9 (.13)
GKG2i12_G1i13		14 (.18)	9 (.28)	10 (.17)	15 (.15)	13 (.15)
<i>GKi13_G1i15</i>		3 (.21)	9 (.33)	6 (.19)	2 (.11)	1 (.06)
<i>GKi14_G1i16_G2i13</i>		3 (.14)	17 (.33)	10 (.17)	8 (.16)	2 (.11)

*Note.*  $n = 986$  valid grade K tests conducted. Italicized items were removed as a result of initial screening. Items that were not answered were recorded as “DNS” Item responses that were unclear were recorded as “UI.”



Table 1.2. Proportion of Grade 1 Student Responses by Item

Item	Item description	Correct response	Most frequent incorrect responses			
		Response (%)	Response (%)	Response (%)	Response (%)	Response (%)
GKi4_G1i1		7 (.96)	6 (.02)	8 (.01)	DNS (.01)	10 (.00)
GKi6_G1i2		9 (.80)	10 (.11)	7 (.03)	1 (.02)	8 (.02)
GKi7_G1i3_G2i1		13 (.69)	15 (.14)	1 (.06)	2 (.05)	0 (.05)
G1i4_G2i2		16 (.59)	11 (.14)	15 (.12)	10 (.07)	7 (.06)
G1i5_G2i3		27 (.24)	16 (.28)	28 (.16)	47 (.15)	36 (.15)
GKi8_G1i6		7 (.78)	1 (.08)	6 (.06)	4 (.04)	3 (.03)
GKi9_G1i7		3 (.72)	2 (.09)	4 (.07)	4 (.06)	9 (.05)
GKi10_G1i8_G2i5		8 (.35)	6 (.24)	2 (.17)	4 (.16)	12 (.06)
G1i9_G2i6		6 (.25)	13 (.38)	20 (.15)	7 (.11)	17 (.09)
G1i10_G2i4		6 (.39)	11 (.21)	16 (.20)	5 (.16)	DNS (.02)
G1i11_G2i7		6 (.23)	9 (.44)	12 (.20)	3 (.06)	27 (.05)
GKG2i11_G1i12		11 (.71)	10 (.13)	7 (.05)	6 (.04)	9 (.04)
GKG2i12_G1i13		14 (.56)	15 (.13)	13 (.12)	9 (.08)	10 (.07)
G1i14_G2i16		26 (.36)	27 (.24)	25 (.13)	16 (.12)	24 (.10)
GKi13_G1i15		3 (.40)	9 (.41)	6 (.08)	DNS (.04)	2 (.04)
GKi14_G1i16_G2i13		3 (.40)	17 (.36)	10 (.07)	DNS (.07)	8 (.06)
G1i17_G2i14		10 (.34)	30 (.33)	11 (.10)	20 (.08)	DNS (.08)
G1i18		6 (.34)	18 (.33)	12 (.10)	5 (.08)	DNS (.08)
G1i19		11 (.37)	19 (.28)	14 (.10)	12 (.09)	DNS (.09)
G1i20		7 (.29)	23 (.30)	8 (.13)	15 (.10)	6 (.10)
G1i21_G2i17		15 (.29)	21 (.24)	16 (.16)	13 (.12)	DNS (.10)

Note.  $n = 1,763$  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	Most frequent incorrect responses			
		Response (%)	Response (%)	Response (%)	Response (%)	Response (%)
GKi7_G1i3_G2i1		13 (.89)	15 (.04)	1 (.02)	0 (.02)	2 (.02)
G1i4_G2i2		16 (.83)	7 (.05)	15 (.05)	11 (.04)	10 (.03)
G1i5_G2i3		27 (.57)	26 (.21)	36 (.09)	28 (.07)	47 (.06)
G1i10_G2i4		6 (.65)	16 (.16)	11 (.11)	5 (.07)	DNS (.01)
GKi10_G1i8_G2i5		8 (.57)	6 (.21)	4 (.10)	2 (.07)	12 (.04)
G1i9_G2i6		6 (.62)	20 (.16)	13 (.10)	7 (.07)	17 (.04)
G1i11_G2i7		6 (.56)	12 (.22)	9 (.18)	3 (.02)	27 (.02)
G2i8		24 (.52)	10 (.27)	16 (.11)	6 (.05)	4 (.03)
G2i9		4 (.47)	3 (.23)	15 (.13)	9 (.10)	12 (.05)
G2i10		5 (.40)	60 (.16)	50 (.16)	10 (.14)	40 (.13)
GKG2i11_G1i12		11 (.93)	10 (.04)	9 (.01)	DNS (.01)	6 (.01)
GKG2i12_G1i13		14 (.86)	15 (.05)	13 (.05)	DNS (.02)	9 (.01)
GKi14_G1i16_G2i13		3 (.82)	17 (.12)	DNS (.02)	7 (.02)	8 (.01)
G1i17_G2i14		10 (.77)	30 (.11)	9 (.05)	11 (.03)	DNS (.03)
G2i15		21 (.75)	20 (.08)	7 (.06)	22 (.05)	17 (.04)
G1i14_G2i16		26 (.70)	27 (.09)	25 (.08)	16 (.05)	24 (.05)
G1i21_G2i17		15 (.63)	16 (.12)	14 (.09)	13 (.06)	DNS (.05)
G2i18		50 (.51)	40 (.12)	41 (.12)	49 (.11)	30 (.08)
G2i19		35 (.30)	45 (.31)	40 (.12)	85 (.11)	36 (.09)
G2i20		2 (.22)	18 (.35)	10 (.12)	11 (.12)	1 (.12)

Note.  $n = 1,737$  valid grade 2 tests conducted. Items that were not answered were recorded as “DNS.” Item responses that were unclear were recorded as “UI.”