

Elementary Mathematics Student Assessment

Measuring the Performance of Grade 3, 4, and 5
Students in Number (Whole Numbers and Fractions),
Operations, and Algebraic Thinking in Fall 2015

Robert C. Schoen
Daniel Anderson
Claire M. Riddell
Charity Bauduin

MAY 2018

Research Report No. 2018-24

SECURE VERSION

The research and development reported here were 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., Riddell, C. M., & Bauduin, C. (2018). *Elementary Mathematics Student Assessment: Measuring the performance of grade 3, 4, and 5 students in number (whole numbers and fractions), operations, and algebraic thinking in fall 2015* (Research Report No. 2018-24). Tallahassee, FL: Learning Systems Institute, Florida State University. DOI: 10.33009/fsu.1581609234.

Copyright 2018, Florida State University. All rights reserved. Requests for permission to use this test should be directed to Robert Schoen, 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).

Elementary Mathematics Student Assessment

Measuring the Performance of Grade 3, 4, and 5 Students in Number (Whole Numbers and Fractions), Operations, and Algebraic Thinking in Fall 2015

Research Report No. 2018-24

Robert C. Schoen¹

Daniel Anderson²

Claire M. Riddell¹

Charity Bauduin¹

May 2018

¹Florida State University

²University 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. Claire Riddell assisted with the development and production of the test and the scoring criteria. 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. She also provided editing and conceptual feedback for the report. Kristy Farina designed and managed the data-entry system, trained data-entry personnel on the system, verified the accuracy of the data, and assisted with description of the data-entry process and sample descriptives for the present report. Alexandra Utecht, Shelby McCrackin, Senai Tazaz, and Claire Riddell served as data-entry personnel. Claire Riddell, Mark McClure, and Monica Hurdal served as reviewers of the test items, response options, and scoring. Xiaotong Yang, Ahmet Guven, and Lanrong Li provided assistance with calculating the ICCs and R^2 estimates based on the field-test data. Anne Thistle provided valuable assistance with copy editing.

We would like to acknowledge the important contributions made by the reviewers of the early drafts of the test and express our gratitude for their contributions of expertise. These reviewers include Charity Bauduin, Wendy Bray, Zachary Champagne, Monica Hurdal, Naomi Iuhasz-Velez, and Mark McClure.

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	xiii
Purpose	xiii
Content	xiii
Sample and Setting	xiv
Test Specifications and Administration.....	xiv
Data Entry and Scoring.....	xiv
Vertical Scaling	xv
Reliability.....	xv
Predictive Validity.....	xv
Summary	xvi
1. Introduction and Overview	1
1.1. Test Overview	2
1.1.1. Fractions on the Number Line Section	3
1.1.2. Parts and Wholes Section	4
1.1.3. Comparing Fractions Section	5
1.1.4. Word Problems Section.....	6
1.1.5. Computation	7
1.1.6. Detailed Test Blueprint.....	8
1.2. Test Administration.....	10
1.3. Description of the Sample and Setting.....	10
2. Test Development, Scoring, and Data-Entry Procedures.....	12
2.1. Content.....	12
2.2. Instrument Development Process	12
2.3. Test Design and Assembly	13
2.4. Test Production	14
2.5. Test Administration for the Fall 2015 EMSA 3–5 Test.....	14
2.6. Data Entry and Verification Procedures	15
2.7. Item-Scoring Procedures	15

3. Data Analysis	16
3.1. Initial Screening With Classical Item Analysis.....	16
3.1.1. Classical Item Difficulty.....	16
3.1.2. Classical Item Discrimination	17
3.1.3. Item/Raw Score Plots	17
3.2. Exploratory Factor Analysis	17
3.3. Specification of Models Based on Item-response Theory.....	20
3.4. Vertical Linking.....	21
3.5. Predictive Validity	21
4. Results.....	22
4.1. Initial Screening of Items.....	22
4.2. Item Response Theory Models	25
4.3. Reliability	30
4.4. Predictive Validity	34
5. Discussion and Reflection	36
References	37

List of Appendices

Appendix A. Grade 3 Test	40
Appendix B. Grade 4 Test.....	49
Appendix C. Grade 5 Test.....	61
Appendix D. Grade 3 Administration Guide.....	73
Appendix E. Grade 4 Administration Guide	82
Appendix F. Grade 5 Administration Guide	91
Appendix G. Scoring Key	100
Appendix H. Results of Initial Screening	105
H.1. Item-level Statistics.....	105
H.2. Spaghetti Plots.....	108
Appendix I. Most Common Incorrect Responses for Each Item	110

List of Tables

Table 1.1. Final Blueprint for the Fall 2015 3–5 EMSA Test	2
Table 1.2. Items in the Fractions on the Number Line Section	4
Table 1.3. Items in the Parts and Wholes Section	5
Table 1.4. Comparing Fractions Section	6
Table 1.5. Word Problems Section	7
Table 1.6. Computation Section	8
Table 1.7. Detailed Test Blueprint for the Fall 2015 3–5 EMSA	9
Table 1.8. Demographic Characteristics of the Students in the Spring 2016 Field Test of the 3–5 EMSA Tests.....	11
Table 2.1. Number of Times the Correct Answer is in Each Position	14
Table 3.1. Number of Factors Suggested by the Minimum Average Partial and Very Simple Structure Tests.....	18
Table 4.1. Item Statistics for Items Removed from Scale during Screening Process.....	22
Table 4.2. Distribution of Item Difficulties and Discrimination Point Estimates for Items Used in the Final Scales	23
Table 4.3. Grade 3 Vertical and Within-Grade-Level Scales IRT Estimates	26
Table 4.4. Grade 4 Vertical and Within-grade-level Scales IRT Estimates	27
Table 4.5. Grade 5 Vertical and Within-grade-level Scales IRT Estimates	28
Table 4.6. Scaling Coefficients Used to Transform the Within-Grade Scales to a Common, Vertical Scale	29
Table 4.7. Sample Descriptives for the Ability Estimates Generated by the Fall 2015 3–5 EMSA and Spring 2016 3–5 EMSA Tests, Split by Grade Level (Students with Both Fall and Spring Scores Only).....	34
Table 4.8. Correlation among Individual Students’ Fall 2015 and Spring 2016 EMSA Test Scores	35
Table 4.9. Intraclass Correlation Coefficients, Disaggregated by Grade Level.....	35
Table G.1. Grade 3 Scoring Key	100
Table G.2. Grade 4 Scoring Key	101
Table G.3. Grade 5 Scoring Key	102
Table H.1. Item Statistics for the Grade 3 Test Based on the Grade 3 Sample (n = 1,045)	105
Table H.2. Item Statistics for the Grade 4 Test Based on the Grade 4 Sample (n = 663)	106
Table H.3. Item Statistics for the Grade 5 Test Based on the Grade 5 Sample (n = 906)	107
Table I.1. Proportion of Grade 3 Student Responses by Item.....	110

Table I.2. Proportion of Grade 4 Student Responses by Item..... 111

Table I.3. Proportion of Grade 5 Student Responses by Item..... 112

List of Figures

Figure 3.1. Parallel analysis scree plot for the grade 3 test.	18
Figure 3.2. Parallel analysis scree plot for the grade 4 test.	19
Figure 3.3. Parallel analysis scree plot for the grade 5 test.	20
Figure 4.1. Distribution of the number of items answered correctly in the final, 16-item scale administered to the grade 3 sample (n = 1,045).	24
Figure 4.2. Distribution of the number of items answered correctly in the final, 29-item scale administered to the grade 4 sample (n = 663).	24
Figure 4.3. Distribution of the number of items answered correctly in the final, 29-item scale administered to the grade 5 sample (n = 906).	25
Figure 4.4. Test characteristic curves for grades 3, 4, and 5 after vertical equating.	29
Figure 4.5. Test information functions for grades 3, 4, and 5.	30
Figure 4.6. Grade 3 item-person plot.	31
Figure 4.7. Grade 4 item-person plot.	32
Figure 4.8. Grade 5 item-person plot.	33
Figure H.1. Grade 3 spaghetti plot.	108
Figure H.2. Grade 4 spaghetti plot.	109
Figure H.3. Grade 5 spaghetti plot.	109

List of Equations

Equation 1. Two-parameter logistic (2PL) item response theory (IRT) model (1)..... 20

List of Abbreviations

2PL.....	Two-parameter logistic
CCSS-M.....	Common Core State Standards for Mathematics
CGI.....	Cognitively Guided Instruction
CR.....	Constructed Response
CTT.....	Classical Test Theory
DNS.....	Did Not Solve
ELL.....	English Language Learner
EMSA.....	Elementary Mathematics Student Assessment
ICC.....	Intraclass Correlation Coefficient
IEP.....	Individualized Education Program
IRT.....	Item Response Theory
MD.....	Measurement Division
MG.....	Multiplication Grouping
SR.....	Selected Response
SWD.....	Students With Disabilities
UI.....	Unclear Intent

Executive Summary

This report describes the Elementary Mathematics Student Assessment (EMSA) as it was used with grade 3, 4, and 5 students in fall 2015. Although the EMSA exists in several versions, each designed for different purposes, we will refer to this specific set of forms as the Fall 2015 3–5 EMSA.

The Fall 2015 3–5 EMSA is designed to serve as a mathematics achievement test administered to students at the beginning of the school year. It includes items involving whole numbers and fractions, number lines, word problems, addition and subtraction involving multidigit whole numbers, and fractions in abstract-symbolic form. The test forms for the Fall 2015 3–5 EMSA differ for each grade level (i.e., third, fourth, fifth).

Purpose

The Fall 2015 EMSA test was intended 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 serve as 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 3–5 EMSA.

Our primary reason for creating this report was for our own reference. The work was so complex, we needed a detailed record of what we did and learned from the experience. A secondary purpose was to allow the research community to scrutinize our research and to provide critical feedback. We hope a tertiary benefit is to provide those conducting similar investigations with the benefit of our experience.

Our intended audience is researchers and evaluators who may be interested in using the instrument in the future. We hope to provide sufficient information that we or others could replicate the administration and scoring of the data.

Content

In general, the test is designed to align with the core content in the number, operations, and fractions domains in the Common Core State Standards for Mathematics (CCSS-M) and the Mathematics Florida Standards at grades 3, 4, and 5 (NGACBP & CCSSO, 2010). The CCSS-M and the Mathematics Florida Standards are similar to one another, but the two sets of curriculum standards are not identical at these grade levels. One difference is the inclusion of content standards related to student understanding of the equals sign and solving equations for an unknown variable in the Mathematics Florida Standards at grade 4. This topic is represented on the Fall 2015 3–5 EMSA tests.

The conceptual framework for the tests are informed by theorized learning progressions in the domain of fractions (Empson & Levi, 2011; Kiearan, 1976; Siegler & Lortie-Forgues, 2015; Siegler & Pyke, 2013; Siegler, Thompson, & Schneider, 2011). The selection of items used on the Fall 2015 3–5 EMSA tests was informed by large-scale field tests of previous versions of the EMSA and assessment items adapted from versions provided in published literature (Baturo, 2004; Beckmann, 2005; Bright, Behr, Post, & Wachsmuth, 1988; Hackenberg, Norton, Wilkins, & Steffe, 2009; Lamon, 2005; Larson, 1980; Lewis & Perry, 2017; Massachusetts Department of Education, 2013; Pothier & Sawada, 1983; Saxe, Diakow, & Gearhart, 2013; Saxe, Kirby, Kang, Le, & Schneider, 2015; Schoen, LaVenía, Bauduin, & Farina, 2016a; 2016b; Schoen, LaVenía, Champagne, Farina, & Tazaz, 2016; Schoen, Liu, Yang, & Paek, 2017).

Sample and Setting

The Fall 2015 3–5 EMSA tests were administered to 2,614 grade 3, 4, and 5 students in 54 schools located in 11 public school districts in Florida during fall 2015.¹ The sample includes 1,045 grade 3 students, 633 grade 4 students, and 906 grade 5 students. The school districts were primarily using *GoMath!* (Dixon, Larson, M., Leiva, & Adams, 2013), a curriculum series designed to be aligned with the Mathematics Florida Standards (Florida Department of Education, 2014), which are very similar to the Common Core State Standards for Mathematics (CCSS-M; NGACBP & CCSSO, 2010). The statewide assessment of student mathematics achievement in grades 3–5 in Florida during spring 2016 was the Florida Standards Assessment (CCSS-M; NGACBP & CCSSO, 2010).

Test Specifications and Administration

The Fall 2015 3–5 EMSA tests included both selected-response and constructed-response test items at each grade level. On selected-response items, students are asked to mark their answer choices by filling in the bubble beneath the answer choice they thought was correct. When previous field-test data were available, selected-response options were based on common responses students at these grade levels have provided in previous versions of these items presented in a constructed-response format. The response options were presented either horizontally or vertically, and most of these item types have five response options. The grade 3 test included 19 items, the grade 4 test 29 items, and the grade 5 test 29 items.

Students were given both selected-response and constructed-response items. On selected-response items students were instructed to fill in the bubble of the correct answer. The Fall 2015 3–5 EMSA was administered as a paper-and-pencil test in a whole-group setting.

Data Entry and Scoring

Research assistants typed student responses from each page for every student directly into a database using FileMaker Pro software (FileMaker Pro, Version 14.1). A double-entry procedure was performed with a random sample of 11% of the tests. More than 99% of the items were entered consistently by the two coders.

Each item on the Fall 2015 EMSA was designed to have a unique, correct solution. Students could nevertheless generate mathematically equivalent responses (e.g., $\frac{6}{10}$, $\frac{3}{5}$) or response that could be interpreted to be a correct response to the item. To accommodate this possibility, an adjudication committee met to review the set of all responses to each item and determine the set of correct answers.

Final scores were determined by means of a two-parameter logistic model based on item-response theory. The scores were mapped onto a single scale according to the Stocking-Lord method for vertical equating (Kolen & Brennan, 2014).

¹The Administration Guides provided in Appendices D, E, and F show 13 school districts. In some of those districts, only grade K–2 teachers participated, and those are not part of this report, but the beginning pages were the same and were used in all grades, K–5.

Vertical Scaling

The large number of items that are common to the grade 3, grade 4, and grade 5 tests allowed for the vertical scaling of the three forms. Vertical scaling opens the possibility for analyses that pool data across the three grade levels. By design, at least three items in each section of the tests were common across adjacent grade levels. Before scaling, all items were investigated relative to classical test theory (CTT) indicators of difficulty (proportion correct) and the potential for the item to discriminate between students of differing abilities (point-biserial correlations). Items with overly low or high difficulty, or with an overly low discrimination index, were evaluated for possible removal before scaling. A two-parameter logistic item response theory (2PL IRT) model was then fit to the data within each grade, which concurrently estimated the item difficulty and discrimination of each item as latent parameters.

Differences between scales were then evaluated on the basis of the common items. Equating constants (the A and B constants; see Kolen & Brennan, 2014) were estimated according to the Stocking-Lord procedure, which focuses on differences between the test characteristic curve of the two test forms. After the constants were estimated, standard equating transformations were applied (Kolen & Brennan, 2014) to transform each of the grade 3 and grade 5 scales to the grade 4 scale. An item difficulty on the vertical scale of 0.00 therefore represents an item for which the average fourth grade student would have a 50% chance of providing a correct response. Items with lower (negative) difficulty are easier; items that are higher (positive) are more difficult. We would expect the average item difficulty in grade 3 to be slightly positive, approximately 0 in grade 4, and slightly negative in grade 5. Student-ability estimates are reported on a similar scale; ability estimates higher than the average fourth grade student are positive; those lower than the average fourth grade student are negative. Because children tend to increase their mathematical abilities over these three grade levels, on average, we would therefore expect the average student ability to be slightly negative in grade 3, approximately 0 in grade 4, and slightly positive in grade 5.

Reliability

Estimates of Cronbach’s alpha exceeded the .80 threshold for a wide range of student abilities in the grade 4 and 5 samples. The range was more restricted in the grade 3 sample, but the reliability estimate was still reasonably high for the test’s intended purpose.

Predictive Validity

The Fall 2015 3–5 EMSA test scores explained a substantial portion of the variance in students’ Spring 2016 3–5 EMSA test scores. For the overall sample, the fall scores explained 56.3% of the variance in the spring scores. In grade 3, the vertically equated scores on the fall test explained 28% of the variance in the vertically equated scores on the spring test; in grade 4, 68%; and in grade 5, 79%. These results support the assertion that the Fall 2015 3–5 EMSA test was reasonably well-suited for its intended use as a baseline covariate for student achievement in the larger study.

The intraclass correlation coefficient (ICC) for the test scores was calculated for the school and class level in a three-level model with students at level one, classes at level two, and schools at level three. The ICCs for the classes were .034, .125, and .186 for grades 3, 4, and 5, respectively. The ICCs for schools were .124, .080, and .059 for grades 3, 4, and 5, respectively.

Summary

The Fall 2015 Grades 3–5 EMSA tests:

- Are designed to serve as a measure of mathematical achievement of grade 3, 4, and 5 students
- Align with the mathematics topics as described by the Mathematics Florida Standards and the Mathematics Common Core State Standards for grades 3, 4, and 5 during the 2015–16 school year in the domains of Operations and Algebraic Thinking, Number and Operations in Base Ten, and Number and Operations—Fractions.
- Focus on early fractions concepts. Fractions concepts involve interpretation of linear representations of fractions, including fractions represented as points on the number line, part-whole relations, problem solving and modeling, and computation.
- Were field tested with a diverse sample of 2,614 grade 3, 4, and 5 students in fall 2015 in 54 schools located in 11 school districts in Florida
- Yield within-grade scores as well as a vertically scaled score to permit direct comparison of scores among students at different grade levels.

The Fall 2015 Grades 3–5 EMSA tests were designed to serve as a baseline measure of student mathematical abilities at the beginnings of grades 3, 4, and 5. The scores were intended for use to investigate baseline equivalence of clusters of students assigned to different treatment conditions and as covariates to control for baseline student ability levels in a randomized-controlled trial of a teacher professional-development program. The tests were not designed to discriminate among individual students or determine cut scores. Further development and standard setting would be required if the scores were to be used for those purposes.

1. Introduction and Overview

The Fall 2015 3–5 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 Mathematics Performance and Cognition interview (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016). Like the Spring 2016 3–5 EMSA, the Fall 2015 3–5 EMSA is scored on a single, vertically equated scale.

The Fall 2015 3–5 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 student achievement in early fractions. 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 measure of student achievement that would be used as a measure of baseline student achievement in a randomized controlled trial of a teacher professional-development program called Cognitively Guided Instruction (CGI).

The Fall 2015 3–5 EMSA has five sections: Fractions on the Number Line, Parts and Wholes, Comparing Fractions, Word Problems, and Computation. Additional information about the composition of each of the five sections is provided in sections 1.1.1–1.1.5 of this report.

The Fall 2015 3–5 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 3–5 EMSA tests were designed to be administered in a whole-group setting in a paper-pencil format. Test administrators were given an administration guide explaining how to administer the tests, along with a script to use while administering them.

The current report focuses on the content, administration protocol, scoring procedures, and psychometric properties for the Fall 2015 3–5 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 3, grade 4, and grade 5 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 3–5 EMSA. The first steps in the analytic process involved initial screening of the test items by means of statistical techniques based on 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 3–5 EMSA Test

Section	Number of items		
	Grade 3	Grade 4	Grade 5
Fractions on the Number Line	3	6	6
Parts and Wholes	3	6	6
Comparing Fractions	6	6	6
Word Problems	5	6	6
Computation	2	5	5
Total	19	29	29

By design, at least three items were identical within each of the four sections on test forms at adjacent grade levels, to permit vertical scaling across grade levels. Only two items are identical on test forms at adjacent grade levels in the Computation section.

All the items on the grade 3 assessment were also present on the grade 4 assessment. All but one of the items on the grade 4 assessment was also on the grade 5 assessment. Generally, the questions that are identical across all three grade levels are presented in the same order. In three instances the order differs in the grade 3 and grade 4 tests, because of a clerical error in the form-creation process. The questions in the anchor set for grades 4 and 5 are presented in the same order on the test.

Test administrators were provided with paper copies of the tests, a class roster, and administration instructions (see Appendices D, E, and F). The administration instructions asked that test administrators adhere to the testing guidelines outlined in the document, which included that the test be administered to the group as a whole but that students complete the assessment independently. In addition, students were asked to write their answers directly in the test booklets. Students were allowed to use blank space provided in the test booklet as scratch paper. In most cases, the students' classroom teacher administered the test.

The testing conditions for the Fall 2015 3–5 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 3–5 EMSA if they were ordinarily permitted to do so in that particular classroom.

Test administrators were also asked to provide students with a comfortable testing environment. The administration instructions deemed it permissible for test administrators to read questions in their

entirety if students struggled to read the problems. Students with special academic plans (e.g., IEP, 504, ELL) were to receive the appropriate testing accommodations as specified in their plans.

Administration of this assessment was estimated to require 45 minutes of class time, but test administrators were instructed not to time it but to allow adequate time to answer the questions.

1.1.1. Fractions on the Number Line Section

This section included questions about students' understanding of the number line as a representational tool for rational numbers. Items in this category included the idea that fractions can be conceptualized as points on the real number line and the idea that the distance from that point to zero tells the magnitude of the number.

These items were presented in constructed-response format. The items displayed number lines, and students were asked to determine what fraction would be represented by a particular point on the number line. Two of the items included fractions with values less than one, and the remaining four items included values greater than one.

Table 1.2. Items in the Fractions on the Number Line Section














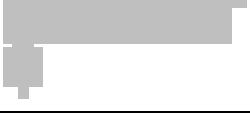
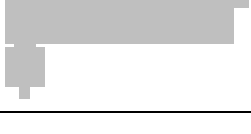
Variable name	Response format	Grade 3	Grade 4	Grade 5
G3G4G5i1b	Constructed			
G3G4G5i2b	Constructed			
G3i5_G4G5i12	Constructed			
G4G5i13a	Constructed			
G4G5i13b	Constructed			
G4G5i13c	Constructed			

Table 1.2 provides an overview of the Fractions on the Number Line items by grade level. The anchor set for grades 3–4 contains three items. The anchor set for grades 4–5 contains six items. Three items in this section were included at all three grade levels, to create a set of anchor items among the three grade levels.

1.1.2. Parts and Wholes Section

This section of the assessment was intended to gather information about students' understanding of fractions as they are conceptualized as a relationship between parts and wholes. The category includes the ideas that (a) fractions are always defined with respect to a referent unit, or whole; (b) fractions consist of a certain number of equal, or same-size, pieces of a whole or unit; and (c) vulgar fractions can be built from the iteration of unit fractions. This category also measures student understanding of the inverse relation between the size of the denominator and the size of the fractional part.

The items in this category were intended to assess students' ability to interpret area representations of fractions. The items also were meant to parse out the interrelated ideas of fractional parts and their referent units. Many of the items we reviewed in the creation of the test showed the shapes already

partitioned for students, but one item requires the students to be the one to do the equal partitioning in order to receive a positive score for the item.

The items in this category were spread throughout the test, because they represent a wide range of difficulty and also have possible connections to other categories, specifically in relation to the number line. As indicated in Table 1.3, three of the six items in this section were given across all three grade levels.

Table 1.3. Items in the Parts and Wholes Section

Variable name	Response format	Grade 3	Grade 4	Grade 5
G3G4G5i3	Constructed			
G3G4G5i4	Constructed			
G4G5i5	Constructed			
G3i8_G4G5i14	Selected			
G4G5i15	Selected			
G4G5i16	Selected			

1.1.3. Comparing Fractions Section

Table 1.4 provides an overview of the Comparing Fractions items by grade level. All six items in this section were included at all three grade levels, to create a set of anchor items among the three grade levels.

Table 1.4. Comparing Fractions Section

Variable name	Response format	Grade 3	Grade 4	Grade 5
G3G4G5i6	Selected			
G3G4G5i7	Selected			
G3i13a_G4G5i18a	Selected			
G3i13b_G4G5i18b	Selected			
G3i13c_G4G5i18c	Selected			
G3i13d_G4G5i18d	Selected			

Note. The three items in boldface font were on the test form but were removed from the final scale at grade 3 as a result of poor item statistics.

1.1.4. Word Problems Section

This section involves the connection of situations described in a narrative that involves solving the problem or matching the situation with equations or other representations to model the situation in which fractions are involved within the narrative, or as a result of the problem.

These items were spread throughout the test, and all but one were posed as open-ended questions with room on the page for students to work on the problem. The test administration protocol specified that if students had trouble reading the problem, the test administrator could read the problem to a student. As indicated in Table 1.5, four items are included in the anchor set across grades 3–5.

Table 1.5. Word Problems Section

Variable name	Response format	Grade 3	Grade 4	Grade 5
G3G4G5i1a	Constructed			
G3G4G5i2a	Constructed			
G3G4i10_G5i9	Constructed			
G4i11_G5i10	Constructed			
G3i11_G4i8	Selected			
G5i11	Constructed			
G3i12_G4G5i17	Selected			

1.1.5. Computation

Table 1.6 shows the computation items across all three grade levels. This section includes computation items with either addition or subtraction. Three of the four items included at least one number presented as a common fraction.

All of these computation items were presented in the middle of the test and were sequenced from least to most difficult. Grade three was only presented with one of the items

Table 1.6. Computation Section

Variable name	Response format	Grade 3	Grade 4	Grade 5
G3G4i9a_G5i8a	Constructed			
G3G4i9b_G5i8b	Constructed			
G4i9c_G5i8c	Constructed			
G4i9d_G5i8d	Constructed			
G4i9e_G5i8e	Constructed			

1.1.6. Detailed Test Blueprint

Table 1.7 provides a detailed blueprint showing the items in each of the five sections of the test (i.e., Fractions on the Number Line, Parts and Wholes, Comparing Fractions, Word Problems, and Computation). Items displayed in strikethrough font were on the test form but were removed from the final scale as a result of poor item statistics. See Chapter 3 of the present report for more information on the review and analysis of the individual items.

Table 1.7. Detailed Test Blueprint for the Fall 2015 3–5 EMSA

Item description	Response format	Variable names		
		Grade 3	Grade 4	Grade 5
<i>Fractions on the Number Line</i>				
[Redacted]	Constructed	G3G4G5i1b	G3G4G5i1b	G3G4G5i1b
[Redacted]	Constructed	G3G4G5i2b	G3G4G5i2b	G3G4G5i2b
[Redacted]	Constructed	G3i5_G4G5i12	G3i5_G4G5i12	G3i5_G4G5i12
[Redacted]	Constructed		G4G5i13a	G4G5i13a
[Redacted]	Constructed		G4G5i13b	G4G5i13b
[Redacted]	Constructed		G4G5i13c	G4G5i13c
<i>Parts and Wholes</i>				
[Redacted]	Constructed	G3G4G5i3	G3G4G5i3	G3G4G5i3
[Redacted]	Constructed	G3G4G5i4	G3G4G5i4	G3G4G5i4
[Redacted]	Constructed		G4G5i5	G4G5i5
[Redacted]	Selected	G3i8_G4G5i14	G3i8_G4G5i14	G3i8_G4G5i14
[Redacted]	Selected		G4G5i15	G4G5i15
[Redacted]	Selected		G4G5i16	G4G5i16
<i>Comparing Fractions</i>				
[Redacted]	Selected	G3G4G5i6	G3G4G5i6	G3G4G5i6
[Redacted]	Selected	G3G4G5i7	G3G4G5i7	G3G4G5i7
[Redacted]	Selected	G3i13a_G4G5i18a	G3i13a_G4G5i18a	G3i13a_G4G5i18a
[Redacted]	Selected	G3i13b_G4G5i18b	G3i13b_G4G5i18b	G3i13b_G4G5i18b
[Redacted]	Selected	G3i13c_G4G5i18c	G3i13c_G4G5i18c	G3i13c_G4G5i18c
[Redacted]	Selected	G3i13d_G4G5i18d	G3i13d_G4G5i18d	G3i13d_G4G5i18d
<i>Word Problems</i>				
[Redacted]	Constructed	G3G4G5i1a	G3G4G5i1a	G3G4G5i1a
[Redacted]	Constructed	G3G4G5i2a	G3G4G5i2a	G3G4G5i2a
[Redacted]	Constructed	G3G4i10_G5i9	G3G4i10_G5i9	G3G4i10_G5i9
[Redacted]	Constructed		G4i11_G5i10	G4i11_G5i10
[Redacted]	Selected	G3i11_G4i8	G3i11_G4i8	
[Redacted]	Constructed			G5i11
[Redacted]	Selected	G3i12_G4G5i17	G3i12_G4G5i17	G3i12_G4G5i17
<i>Computation</i>				
[Redacted]	Constructed	G3G4i9a_G5i8a	G3G4i9a_G5i8a	G3G4i9a_G5i8a
[Redacted]	Constructed	G3G4i9b_G5i8b	G3G4i9b_G5i8b	G3G4i9b_G5i8b
[Redacted]	Constructed		G4i9c_G5i8c	G4i9c_G5i8c
[Redacted]	Constructed		G4i9d_G5i8d	G4i9d_G5i8d
[Redacted]	Constructed		G4i9e_G5i8e	G4i9e_G5i8e
Items on test form		19	29	29
Items in final scale		16	29	29

Note. The three items in strikethrough font were on the test form but were removed from the final scale at grade 3 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. 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 3–5. Tests forms were delivered to schools by project staff during the week of preplanning (i.e., the week before students returned to school for the year). In the field tests reported here, the students' classroom teachers administered the tests during the first two weeks of the school year in all except five classrooms.

The analytical samples included 2,614 students, 1,045 in grade 3, 633 in grade 4, and 906 in grade 5. These students represented 266 classrooms in 11 Florida public school districts. Table 1.8 provides descriptive statistics for the data available at the time of this report.

Table 1.8. Demographic Characteristics of the Students in the Spring 2016 Field Test of the 3–5 EMSA Tests

Student characteristic	Number (proportion of sample or subsample)			
	Grade 3 (<i>n</i> = 1,045)	Grade 4 (<i>n</i> = 663)	Grade 5 (<i>n</i> = 906)	Overall sample (<i>n</i> = 2,614)
Gender				
Male	157 (.15)	58 (.09)	128 (.14)	343 (.13)
Female	162 (.16)	60 (.09)	134 (.15)	356 (.14)
Unknown	726 (.69)	545 (.82)	644 (.71)	1,915 (.73)
Language				
ELL	2 (<.01)	1 (<.01)	4 (<.01)	7 (<.01)
Non-ELL	313 (.30)	112 (.17)	252 (.28)	677 (.26)
Unknown	730 (.70)	550 (.83)	650 (.72)	1,930 (.74)
Exceptionality				
SWD	35 (.03)	13 (.02)	33 (.04)	81 (.03)
Non-SWD	276 (.26)	102 (.15)	185 (.20)	563 (.22)
Gifted	8 (.01)	3 (<.01)	44 (.05)	55 (.02)
Unknown	726 (.69)	545 (.82)	644 (.71)	1,915 (.73)
Race				
White	161 (.15)	51 (.08)	104 (.11)	316 (.12)
Black	22 (.02)	15 (.02)	24 (.03)	61 (.02)
Asian	8 (.01)	2 (<.01)	4 (<.01)	14 (.01)
Other	28 (.03)	14 (.02)	20 (.02)	62 (.02)
Unknown	826 (.79)	581 (.88)	754 (.83)	2,161 (.83)
Ethnicity				
Hispanic	9 (.01)	3 (<.01)	17 (.02)	29 (.01)
Non-Hispanic	219 (.21)	82 (.12)	152 (.17)	453 (.17)
Unknown	817 (.78)	578 (.87)	737 (.81)	2,132 (.82)

Note. Other individual student demographic characteristics, such as ethnicity, exceptionality, or eligibility for free or reduced-price lunch, were not available at the time the present report was written. Some of the percentages do not sum to 1.00 because of rounding error. SWD = students with disabilities.

2. Test Development, Scoring, and Data-Entry Procedures

2.1. Content

In general, the test was designed to align with the core content in the Number and Base Ten, Number-Fractions, and Operations and Algebraic Thinking domains in the CCSS-M and the Mathematics Florida Standards at grades 3, 4, and 5 (NGACBP & CCSO, 2010; Florida Department of Education, 2014). The conceptual framework for the tests are informed by theorized learning progressions in the domain of fractions (Empson & Levi, 2011; Kiearan, 1976; Siegler & Lortie-Forgues, 2015; Siegler & Pyke, 2013; Siegler, Thompson, & Schneider, 2011). The selection of items used on the Fall 2015 3–5 EMSA tests was informed by large-scale field tests of previous versions of the EMSA and assessment items adapted from versions provided in published literature (Baturo, 2004; Beckmann, 2005; Bright, Behr, Post, & Wachsmuth, 1988; Hackenberg, Norton, Wilkins, & Steffe, 2009; Lamon, 2005; Larson, 1980; Lewis & Perry, 2017; Massachusetts Department of Education, 2013; Pothier & Sawada, 1983; Saxe, Diakow, & Gearhart, 2013; Saxe, Kirby, Kang, Le, & Schneider, 2015; Schoen, LaVenía, Bauduin, & Farina, 2016a; 2016b; Schoen, LaVenía, Champagne, Farina, & Tazaz, 2016; Schoen, Liu, Yang, & Paek, 2017).

The Mathematics Florida Standards drive the accountability system in place in the schools where the field study was conducted. The CCSS-M and the Mathematics Florida Standards are similar to one another, but the two are not identical at these grade levels. One difference is the inclusion of content standards related to student understanding of the equals sign and solving equations for an unknown variable in the Mathematics Florida Standards at grade 4.

Conceptual categories for the test blueprint were determined on the basis of a review of scholarly literature and current standards related to student ability in the domain of problem solving, equality, fractions, and computation. From these sources, five major categories were developed (1) Fractions on a Number Line, Partitioning, and Iterating; (2) Parts and Wholes; (3) Comparing Fractions; (4) Word Problems; and (5) Computation. These categories were determined to be likely to provide important information about the effect of the CGI 3–5 program on students' problem-solving abilities.

2.2. Instrument Development Process

The development process for the Fall 2015 3–5 EMSA tests consisted of the following phases:

1. Review of content expectations for grades 3, 4, and 5 in the CCSS-M (NGACBP & CCSO, 2010) and Mathematics Florida Standards (Florida Department of Education, 2014)
2. Review of the literature and evaluation of the goals of the CGI 3–5 program
3. Development of the first written draft of the test blueprint
4. Review of the draft blueprint by internal members of the evaluation team and external experts in mathematics and mathematics education
5. Revision of the blueprint based on feedback
6. Development of the first written draft of the test form for grades 3, 4, and 5 and corresponding scoring procedures
7. Review of the draft test forms, editing, and proofing
8. Analysis of the frequency of correct response position and distribution of correct response positions across each grade level test
9. Development of administration instructions
10. Proofreading of test and administration instruction forms

Because the test was used in fall 2015 for the purpose of evaluating the impact of a teacher professional-development program based on CGI and designed for grade 3–5 teachers, the corpus of literature related specifically to CGI (Carpenter et al., 2015; Carpenter, Franke, & Levi, 2003; Empson & Levi, 2011) was reviewed. In addition to review and analysis of these published sources, CGI experts were consulted about those aspects of student mathematical ability likely to be affected by a teacher's involvement in the program. To avoid overalignment of the interview with the professional-development program, we took abundant caution to avoid using problems that were encountered by teachers in the professional-development activities. The workshop leaders and coordinators did not have access to the items on the test, and we do not expect that the grade 3, 4, or 5 students who took the test had seen any of these problems.

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. The original draft test was shared with senior project personnel and revised according to internal feedback. Feedback from external experts resulted in changes to items, including types of problems included, numbers used in the problems, administration instructions, and the number of items in each category.

2.3. Test Design and Assembly

Plenty of empty space was available on the page for students to draw or record their thoughts as necessary. The Computation section consisted of items presented as open equations. Each problem is presented as a single equation involving either the addition or subtraction operator and the numbers. Each was presented in the standard (i.e., $a + b = c$, $a - b = c$) form (Stigler, Fuson, Ham, & Kim, 1986; Schoen, Champagne, Whitacre, & McCrackin, in review) with an open box for the missing number. Students write the number that completes the equation in the box to indicate their responses.

For all sections except computation and comparing fractions, no more than three items are presented per page. Large (16-point) Cambria font was used on the computation section. Medium (13-point) Cambria font was used on the other sections. Copies of the grade 3, 4, and 5 tests are presented in Appendices A, B, and C, respectively.

Items on the Fall 2015 3–5 EMSA tests were presented in either a selected-response or a constructed-response format. On selected-response items, five response options were presented horizontally across the page and included exactly one correct response for each item. The response options were always numbers. The students were directed to fill in the circles (which we call bubbles) below or beside their answer choices. Generally, bubbles were centered beneath the corresponding response option, and responses are centered horizontally across the page. In two instances, the bubbles were aligned vertically with the answer choices beside the bubbles. 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 was 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
3	4	3	0	1	1
4	4	4	0	2	1
5	4	4	0	1	1

Note. Five questions presented answer choices only in the A and B positions at each respective grade level.

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

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/constructed-response versions of the items on the Fall 2013 and Fall 2014 grades 1–2 EMSA tests and in the 2014 and 2015 Mathematics Performance and Cognition interviews, as well as other sources.

Test administrators were given permission to read each math problem aloud to students if individual students have difficulty reading the items. 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 was expected to provide any and all required accommodations for those individual students and to document the accommodation on the student information sheet. The test was not designed to be timed, so test administrators were instructed to allow students adequate time to answer all of the questions.

2.4. Test Production

The tests were printed double-sided on 28-pound white paper at Florida State University and distributed to the participating schools. The heavy paper was used, because the optical scanner yields better 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 provided 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 EMSA 3–5 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 teachers 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 3–5 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 3, 4 and 5 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 forms. 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

Research assistants typed student responses into forms hosted on a FileMaker Pro database (FileMaker Pro, Version 14.1). Response fields for multiple choice items allowed only the codes for offered responses, as well as codes for missing or uninterpretable item-level data (UI for unclear intent, DNS for did not solve). The code MR, for multiple response, was used when students attempted to select more than one of the options presented. Most constructed-response items were entered as the student wrote them, and research assistants chose UI or DNS when applicable. The research assistants' task was to interpret both the student's handwriting and the student's intent, with the goal of entering the student's intended response exactly as it was written.

If a student responded to a selected response item with a fraction that was equivalent to one of the response options given, the student's response was coded as though the student had bubbled the corresponding response. When the response given did not match any selected-response option provided, the code CR (for constructed response) was used.

2.7. Item-Scoring Procedures

Constructed-response items were entered as numeric responses, and the full set of responses was scored by an adjudication committee. The set of observed responses that were judged to be correct were provided in the scoring guide. (See Appendix G.) Selected-response items were scored according to the scoring guide. Constructed-response items requiring students to mark their answers on a number line were scored as correct or incorrect by means of an overlay sheet, which provided guidelines for error tolerance. The overlay sheet is also provided in Appendix G.

3. Data Analysis

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 3–5 EMSA were subjected to the following analyses:

1. Initial screening of items by means of classical test theory (CTT)
2. Exploratory Factor Analysis (by means of tetrachoric correlations to avoid arriving upon difficulty-related factors)
3. Within-grade scaling according to a two-parameter logistic item-response theory (2PL IRT) model
4. Equating of scales between grades by means of a nonequivalent groups with anchor tests design (i.e., common items between grades) to create the vertical scale. The Stocking-Lord method (Kolen & Brennan, 2014) was used to transform the within-grade scales to the common, vertical scale.
5. Examination of the ability of Fall 2015 3–5 EMSA scores to predict students' Spring 2016 3–5 EMSA scores²

Initial item screening with CTT identified items that might not be providing useful information about test-takers' abilities (e.g., overly difficult or easy items). Factor analysis tested the dimensionality of the test as a means of determining whether the test was measuring a sufficiently unidimensional construct (see Anderson, Kahn, & Tindal, 2017). This analysis revealed 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.

All analyses and displays of data were conducted within the R statistical computing environment (R Core Team, 2017).

3.1. Initial Screening With Classical Item Analysis

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

3.1.1. Classical Item Difficulty

Each individual item on the Spring 2015 3–5 EMSA was scored dichotomously. For these items, the CTT-based item difficulty statistic, or p -value, corresponded to the proportion of test takers in the within-grade-level samples who produced correct answers 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 might not be contributing useful information to the overall score, but that was not

²Fall 2015 and Spring 2016 EMSA tests were not equated with one another

always the case. At times, those high- or low-difficulty items can be useful for discriminating among test-takers in the corresponding ability range (i.e., very high or low achievement levels). Items scoring below/above these thresholds were more closely examined.

3.1.2. Classical Item Discrimination

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

3.1.3. Item/Raw Score Plots

Additional screening involved the generation of item/raw score plots, where students' total scores were plotted along the horizontal axis, and the proportion responding correctly was mapped onto the vertical axis. Separate lines were produced for each item. (See Appendix H.) Because the sample size for each individual raw score was relatively low, we smoothed the overall relation using local scatterplot smoothing, 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 3–5, such that scores on the grade 3, grade 4, and grade 5 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 (Velicer, 1976), Revelle's very simple structure test (Revelle & Rocklin, 1979), and parallel analysis (Horn, 1965). In cases where these three tests provided conflicting evidence, 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 Minimum Average Partial and Very Simple Structure Tests

Grade level	Minimum average partial	Very simple structure 1	Very simple structure 2
3	2	2	3
4	2	1	2
5	2	1	2

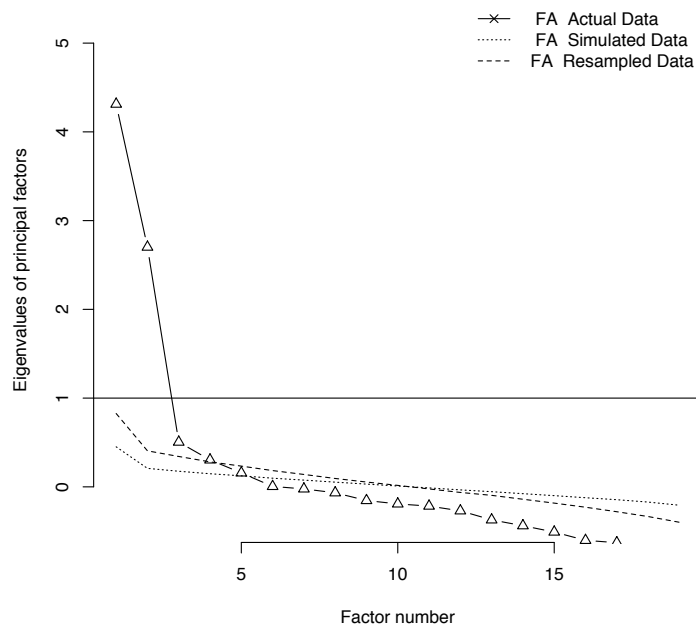


Figure 3.1. Parallel analysis scree plot for the grade 3 test.

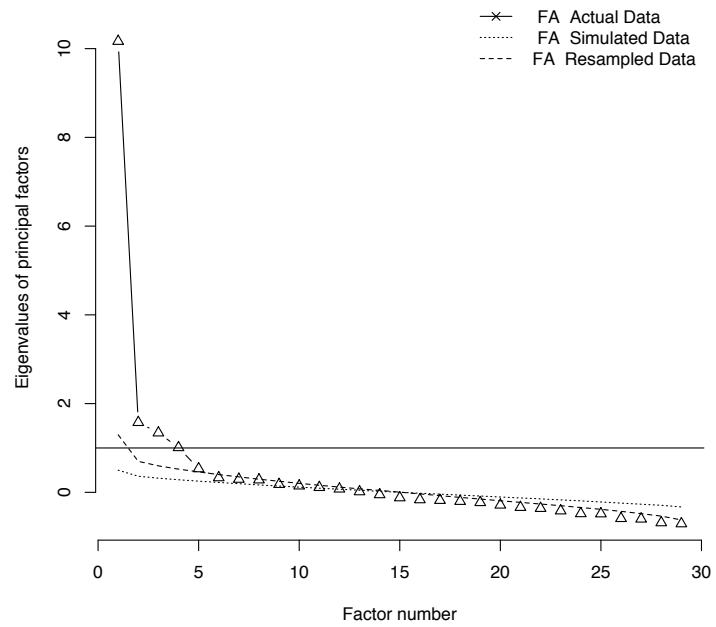


Figure 3.2. Parallel analysis scree plot for the grade 4 test.

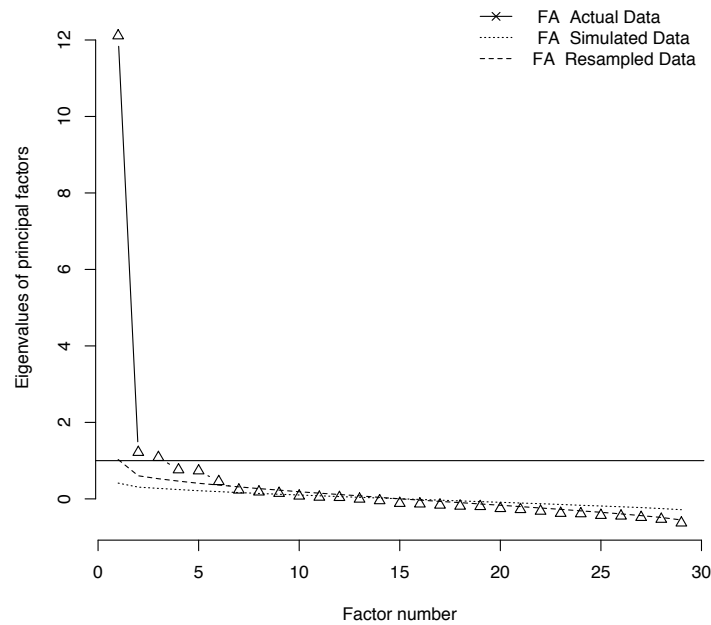


Figure 3.3. Parallel analysis scree plot for the grade 5 test.

Across tests, the optimal number of dimensions to extract was mixed. The Minimum Average Partial test indicated two dimensions across grades, the Very Simple Structure 1 test indicated two dimensions in grade 3 and a unidimensional structure in grades 4 and 5, and the Very Simple Structure 2 test indicated three dimensions at grade 3 and 2 at grades 4 and 5. The scree plots 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 that from the second dimension of the randomly generated data (i.e., parallel analysis always indicated more than one dimension). Collectively, these results indicated that, although more than one dimension was probably present in the data, they were reasonably represented by a single dimension 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 that the purpose of the scaling was to create a single scale across grades 3–5, we proceeded to IRT scaling by assuming a unidimensional structure.

3.3. Specification of Models Based on Item-response Theory

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

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

where θ_j represents the estimated ability of student j , and a_i and b_i are the discrimination and difficulty of item i , respectively. In essence, the log odds of a student's correctly responding to an item are driven by the difference between the student's estimated ability, θ_j , and the difficulty of the item b_i . Log odds

are estimated as the ratio between the odds of a correct 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 θ increases). Items with lower discrimination values are weighted less in the estimation of θ than those with higher values, as the difference between the item difficulty and the students' ability is multiplied by the estimated discrimination of the item.

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 4—the middle of the grade span—and equated both the grade 3 and grade 5 test parameters relative to the grade 4 scale. Because all grade-level test forms included common items, multiple links joined each test and the grade 4 scale. That is, the grade 3 test included a direct link of common items between grades 3 and 4, but also an indirect link through the common items with grade 5. Similarly, grade 5 included both a direct and an indirect link with grade 4. 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 incorporating 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 3 and 5 onto the grade 4 scale by means of standard transformation procedures (see Kolen & Brennan, 2014).

3.5. Predictive Validity

The ability estimates generated with the Fall 2015 3–5 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 3–5 EMSA. This information can provide some evidence of external validity (Flake, Pek, & Hehman, 2017), and it is also useful in estimation of statistical power.




These analyses involved first saving the scale scores from the final, vertically scaled scores for the grade 3, 4, and 5 tests. Then, as manifest variables, the scale scores were merged into a file containing similar scores for the spring 2016 EMSA tests for grades 3–5 (Schoen, Anderson, & Bauduin, 2017). We investigated evidence of predictive validity using a single-level regression model in which the Fall 2015 3–5 EMSA scores predicted the Spring 2016 3–5 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. These analyses were done for the aggregated sample and for the individual grade levels.

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 3, 4, and 5 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. Figures 4.1–4.3 show the raw-score distributions for students on the total test.

Table 4.1. Item Statistics for Items Removed from Scale during Screening Process

Item	Item description	Grade level	CTT-based statistics		Vertically scaled IRT-based statistics	
			PC (se)	PB	Discrim (se)	Diff (se)
G3i13b_G4G5i18b		3	.19 (.012)	.24	–	–
G3i13c_G4G5i18c		3	.17 (.012)	.09	–	–
G3i13d_G4G5i18d		3	.17 (.012)	.28	–	–

Note. CTT = classical test theory; IRT = item response theory; PC = proportion correct; PB = point biserial; Diff = Difficulty; Discrim = discrimination.

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

Value	Number of items		
	Grade 3	Grade 4	Grade 5
	<i>P-value</i>		
>.90	0	0	1
.80–.89	0	0	7
.70–.79	1	3	8
.60–.69	0	1	3
.50–.59	2	9	3
.40–.49	2	5	1
.30–.39	2	3	5
.20–.29	6	4	1
.10–.19	3	2	0
<.09	0	2	0
Mean	0.31	0.43	0.65
Median	0.28	0.44	0.73
Standard	0.17	0.19	0.20
	<i>Point-biserial correlation</i>		
.80–1.0	0	0	0
.60–.79	1	8	10
.40–.59	12	19	17
.20–.39	3	2	2
0.0–.20	0	0	0
Mean	0.40	0.47	0.51
Median	0.42	0.47	0.52
Standard	0.12	0.10	0.10

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

The distribution of raw scores (i.e. number of items answered correctly) for the final set of items in the grade 3, 4, and 5 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 3 test was 16. The raw-score distribution for the grade 3 sample appears to be positively skewed. Almost 2% of the grade 3 sample received a zero score, whereas less than 0.5% of the grade 3 sample received a perfect score. The grade 4 distribution is more symmetric; very few students received a zero score, and 0.5% of students in the sample received a perfect score. The grade 5 sample distribution is negatively skewed. Very few of the students in the grade 5 sample received a zero score, but slightly more than 3% received a perfect score.

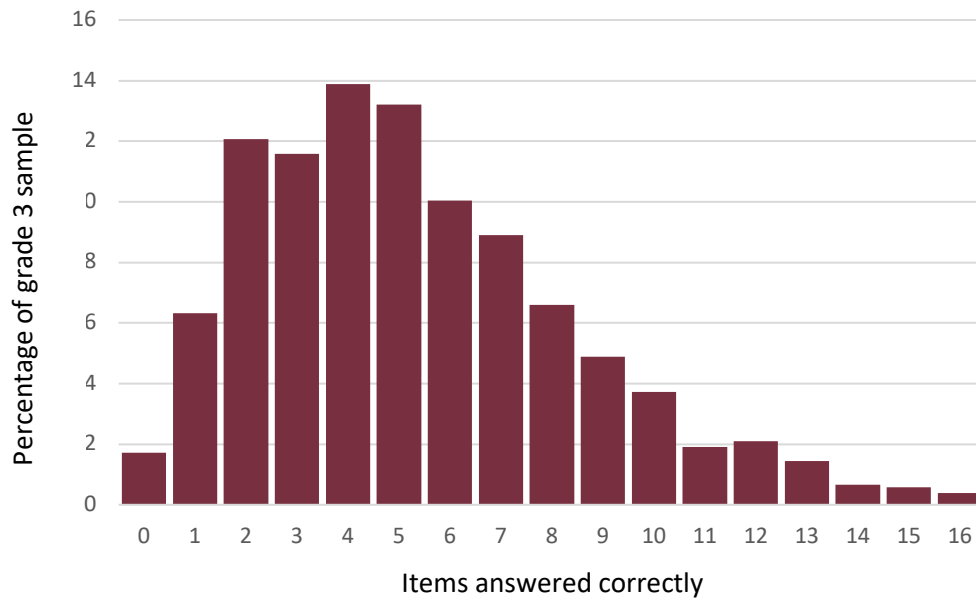


Figure 4.1. Distribution of the number of items answered correctly in the final, 16-item scale administered to the grade 3 sample ($n = 1,045$).

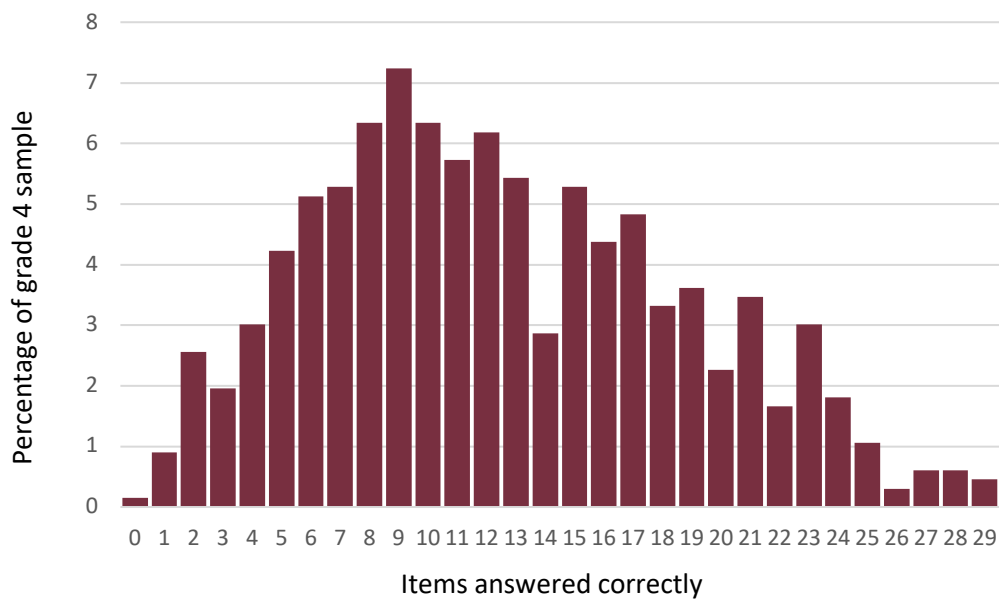


Figure 4.2. Distribution of the number of items answered correctly in the final, 29-item scale administered to the grade 4 sample ($n = 663$).

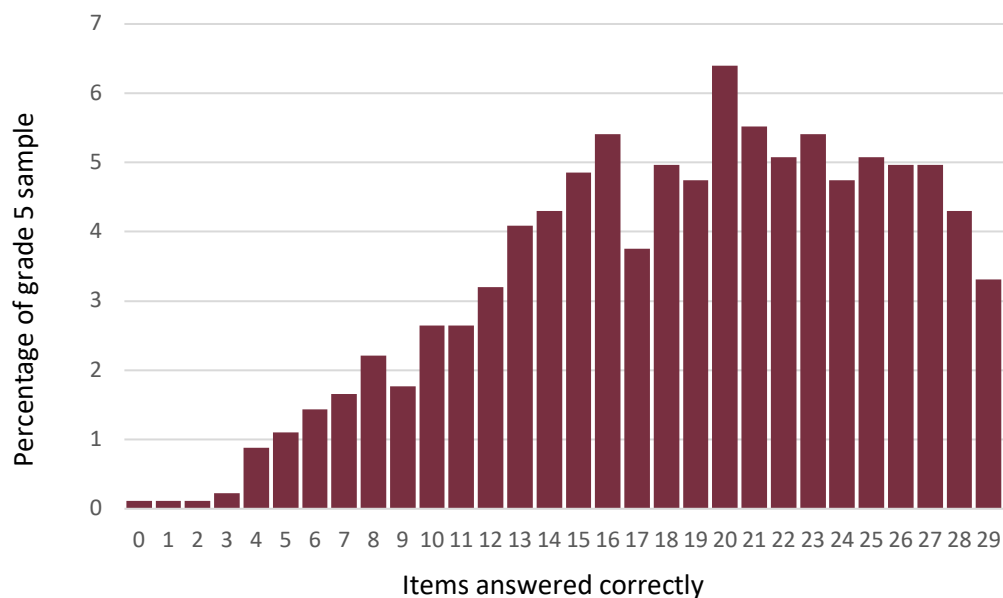


Figure 4.3. Distribution of the number of items answered correctly in the final, 29-item scale administered to the grade 5 sample ($n = 906$).

4.2. Item Response Theory Models

The IRT-based discrimination and difficulty estimates for grades 3, 4, and 5 are presented in Tables 4.3, 4.4, and 4.5, respectively. The mean item difficulties based on the within-grade models were 0.97, 0.13, and -0.95 for grades 3, 4, and 5, respectively. Item discriminations ranged from 0.38 to 5.24 in grade 3, 0.46 to 2.41 in grade 4, and 0.83 to 2.73 in grade 5.

Table 4.3. Grade 3 Vertical and Within-Grade-Level Scales IRT Estimates

Item	Item description	Vertical scale		Within-grade-level scale	
		Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
G3G4G5i1a		5.79 (.524)	−0.82 (.178)	4.21 (.524)	−0.40 (.178)
G3G4G5i1b		7.20 (.843)	−0.66 (.226)	5.24 (.843)	0.64 (.226)
G3G4G5i2a		2.57 (.204)	0.49 (.224)	1.87 (.204)	3.19 (.224)
G3G4G5i2b		2.54 (.170)	−0.08 (.132)	1.85 (.170)	1.72 (.132)
G3G4G5i3		0.94 (.085)	−0.65 (.068)	0.68 (.085)	0.10 (.068)
G3G4G5i4		1.41 (.120)	0.69 (.117)	1.02 (.120)	2.02 (.117)
G3i5_G4G5i12		0.90 (.094)	0.80 (.086)	0.65 (.094)	1.39 (.086)
G3G4G5i6		0.82 (.082)	−0.83 (.067)	0.60 (.082)	−0.06 (.067)
G3G4G5i7		0.94 (.113)	1.55 (.113)	0.68 (.113)	2.16 (.113)
G3i8_G4G5i14		0.95 (.090)	0.34 (.078)	0.69 (.090)	1.04 (.078)
G3G4i9a_G5i8a		1.09 (.094)	0.15 (.079)	0.80 (.094)	0.98 (.079)
G3G4i9b_G5i8b		1.22 (.098)	−0.01 (.080)	0.89 (.098)	0.90 (.080)
G3G4i10_G5i9		1.32 (.101)	0.01 (.084)	0.96 (.101)	1.01 (.084)
G3i11_G4i8		0.61 (.088)	1.45 (.080)	0.45 (.088)	1.35 (.080)
G3i12_G4G5i17		0.60 (.078)	0.12 (.070)	0.44 (.078)	0.52 (.067)
G3i13a_G4G5i18a		0.53 (.086)	−2.75 (.086)	0.38 (.086)	−1.06 (.074)

Table 4.4. Grade 4 Vertical and Within-grade-level Scales IRT Estimates

Item	Item description	Vertical scale		Within-grade-level scale	
		Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
G3G4G5i1a		1.30 (.148)	-0.74 (.116)	1.30 (.148)	-0.97 (.116)
G3G4G5i1b		1.16 (.132)	-0.42 (.102)	1.16 (.132)	-0.49 (.102)
G3G4G5i2a		1.85 (.182)	0.45 (.130)	1.85 (.182)	0.82 (.130)
G3G4G5i2b		1.19 (.130)	0.49 (.103)	1.19 (.130)	0.58 (.103)
G3G4G5i3		1.24 (.154)	-1.27 (.134)	1.24 (.154)	-1.57 (.134)
G3G4G5i4		1.98 (.201)	-0.19 (.129)	1.98 (.201)	-0.38 (.129)
G4G5i5		1.79 (.179)	-0.07 (.120)	1.79 (.179)	-0.12 (.120)
G3G4G5i6		0.82 (.117)	-1.26 (.102)	0.82 (.117)	-1.03 (.102)
G3G4G5i7		2.24 (.224)	0.05 (.136)	2.24 (.224)	0.11 (.136)
G3i11_G4i8		0.75 (.103)	0.55 (.089)	0.75 (.103)	0.41 (.089)
G3G4i9a_G5i8a		0.82 (.107)	-0.03 (.089)	0.82 (.107)	-0.02 (.089)
G3G4i9b_G5i8b		1.28 (.135)	-0.03 (.102)	1.28 (.135)	-0.04 (.102)
G4i9c_G5i8c		1.83 (.185)	0.76 (.147)	1.83 (.185)	1.39 (.147)
G4i9d_G5i8d		1.20 (.131)	0.48 (.104)	1.20 (.131)	0.57 (.104)
G4i9e_G5i8e		2.41 (.386)	2.05 (4.94)	2.41 (.386)	4.94 (.559)
G3G4i10_G5i9		1.02 (.118)	0.43 (.097)	1.02 (.118)	0.44 (.097)
G4i11_G5i10		0.93 (.117)	1.14 (.105)	0.93 (.117)	1.06 (.105)
G3i5_G4G5i12		0.83 (.113)	1.32 (.103)	0.83 (.113)	1.10 (.103)
G4G5i13a		1.59 (.101)	1.77 (.091)	1.59 (.199)	2.80 (.220)
G4G5i13b		0.82 (.173)	-0.14 (.115)	0.82 (.107)	-0.11 (.089)
G4G5i13c		2.52 (.121)	1.93 (.096)	2.52 (.387)	4.87 (.544)
G3i8_G4G5i14		1.55 (.157)	-0.13 (.112)	1.55 (.157)	-0.21 (.112)
G4G5i15		0.57 (.112)	2.64 (.108)	0.57 (.112)	1.50 (.108)
G4G5i16		1.41 (.144)	0.90 (.126)	1.41 (.149)	1.27 (.126)
G3i12_G4G5i17		0.70 (.101)	0.38 (.087)	0.70 (.101)	0.27 (.087)
G3i13a_G4G5i18a		0.46 (.101)	-2.08 (.091)	0.46 (.101)	-0.95 (.091)
G3i13b_G4G5i18b		1.66 (.173)	-0.09 (.115)	1.66 (.173)	-0.14 (.115)
G3i13c_G4G5i18c		1.01 (.121)	0.41 (.096)	1.01 (.121)	0.42 (.096)
G3i13d_G4G5i18d		1.57 (.165)	-0.07 (.112)	1.57 (.165)	-0.10 (.112)

Table 4.5. Grade 5 Vertical and Within-grade-level Scales IRT Estimates

Item	Item description	Vertical scale		Within-grade-level scale	
		Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
G3G4G5i1a		1.46 (.157)	-0.34 (.146)	1.62 (.157)	-2.03 (.146)
G3G4G5i1b		1.18 (.123)	0.50 (.104)	1.31 (.123)	-1.18 (.104)
G3G4G5i2a		1.71 (.156)	0.52 (.116)	1.89 (.156)	-0.91 (.116)
G3G4G5i2b		1.09 (.108)	0.74 (.087)	1.21 (.108)	-0.33 (.087)
G3G4G5i3		1.02 (.165)	-1.78 (.182)	1.14 (.165)	-2.90 (.182)
G3G4G5i4		1.96 (.206)	-0.23 (.194)	2.17 (.206)	-2.51 (.194)
G4G5i5		1.44 (.149)	-0.19 (.133)	1.59 (.149)	-1.78 (.134)
G3G4G5i6		0.81 (.140)	-2.01 (.146)	0.90 (.140)	-2.49 (.146)
G3G4G5i7		2.46 (.271)	-0.18 (.257)	2.73 (.271)	-3.04 (.257)
G3G4i9a_G5i8a		1.15 (.123)	-0.14 (.108)	1.28 (.123)	-1.38 (.108)
G3G4i9b_G5i8b		1.40 (.141)	-0.03 (.122)	1.55 (.141)	-1.51 (.122)
G4i9c_G5i8c		1.69 (.189)	-0.51 (.190)	1.87 (.189)	-2.63 (.190)
G4i9d_G5i8d		0.94 (.125)	-1.02 (.121)	1.04 (.125)	-1.93 (.121)
G4i9e_G5i8e		2.22 (.207)	1.58 (.142)	2.47 (.207)	1.18 (.142)
G3G4i10_G5i9		1.19 (.115)	0.66 (.091)	1.32 (.115)	-0.46 (.091)
G4i11_G5i10		1.31 (.122)	0.78 (.094)	1.46 (.122)	-0.35 (.094)
G5i11		2.44 (.237)	1.70 (.168)	2.70 (.237)	1.60 (.168)
G3i5_G4G5i12		0.75 (.090)	1.70 (.079)	0.83 (.090)	0.48 (.079)
G4G5i13a		1.81 (.164)	1.58 (.120)	2.01 (.164)	0.95 (.120)
G4G5i13b		1.09 (.120)	-0.22 (.107)	1.21 (.120)	-1.39 (.107)
G4G5i13c		2.34 (.246)	2.09 (.210)	2.59 (.246)	2.42 (.210)
G3i8_G4G5i14		1.53 (.154)	-0.08 (.135)	1.70 (.154)	-1.73 (.135)
G4G5i15		0.81 (.094)	1.93 (.083)	0.90 (.094)	0.72 (.083)
G4G5i16		1.38 (.126)	1.13 (.095)	1.53 (.126)	0.11 (.095)
G3i12_G4G5i17		1.01 (.106)	0.43 (.088)	1.12 (.106)	-0.63 (.088)
G3i13a_G4G5i18a		0.90 (.120)	-0.95 (.115)	1.00 (.120)	-1.80 (.115)
G3i13b_G4G5i18b		2.12 (.168)	0.16 (.209)	2.35 (.209)	-1.89 (.168)
G3i13c_G4G5i18c		1.07 (.091)	0.38 (.110)	1.19 (.110)	-0.72 (.091)
G3i13d_G4G5i18d		1.48 (.123)	0.08 (.146)	1.64 (.146)	-1.43 (.123)

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

From	To	A (SE)	B (SE)
3	4	0.74 (0.04)	-0.83 (0.06)
5	4	1.17 (0.05)	1.20 (0.07)

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 inflection points indicate that, as would be expected, the grade 3 test was the easiest. Grades 4 and 5, however, were essentially equivalently difficult. The inflection points for the grade 4 and grade 5 tests were nearly indistinguishable. They can be interpreted as the estimated ability level associated with having a 50% chance of responding to approximately half the items correctly.

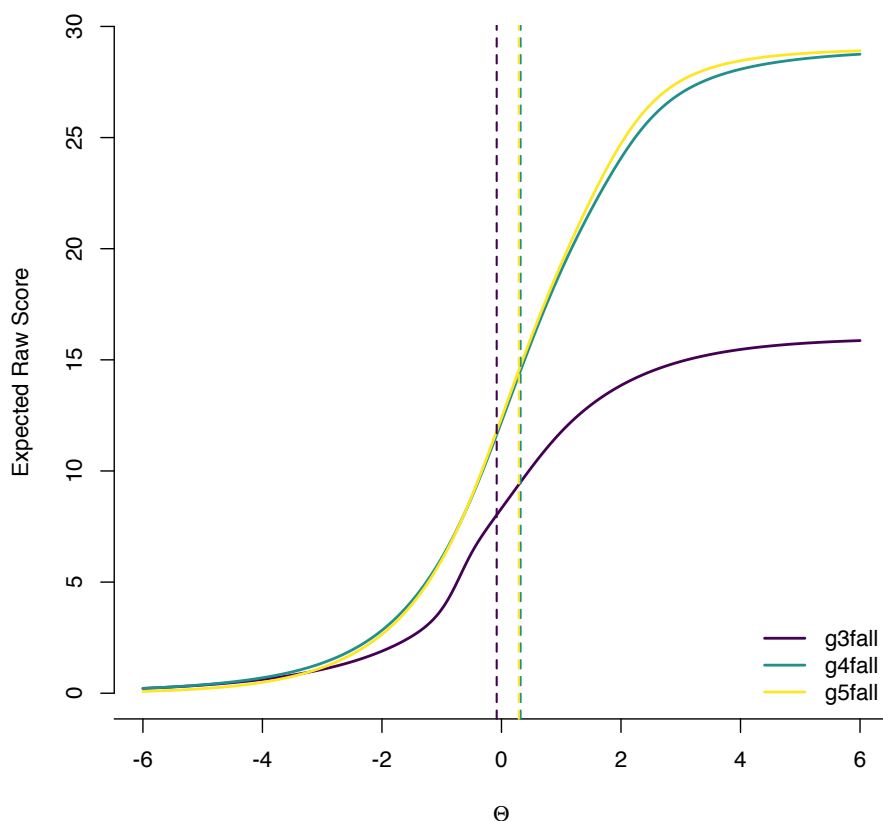


Figure 4.4. Test characteristic curves for grades 3, 4, and 5 after vertical equating.

4.3. Reliability

Item response theory provides a conditional view of reliability, in which the reliability of the measure is viewed as depending on 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 below displays the test information functions for each of the three tests. These functions are test-level summaries of the reliability, each mapped on the common, vertical scale. Under the standardized θ , reliability is equivalent to a Cronbach's alpha of 0.80 when information is equal to 5.0. Therefore, in the figure below, vertical dashed lines display the ability regions for each test in which information is greater than or equal to 5.0 (implying the ranges in which reliability is ≥ 0.80). Notice that grade 3 had a much narrower range in which reliability was equivalent to 0.80 or above than did grades 4 and 5.

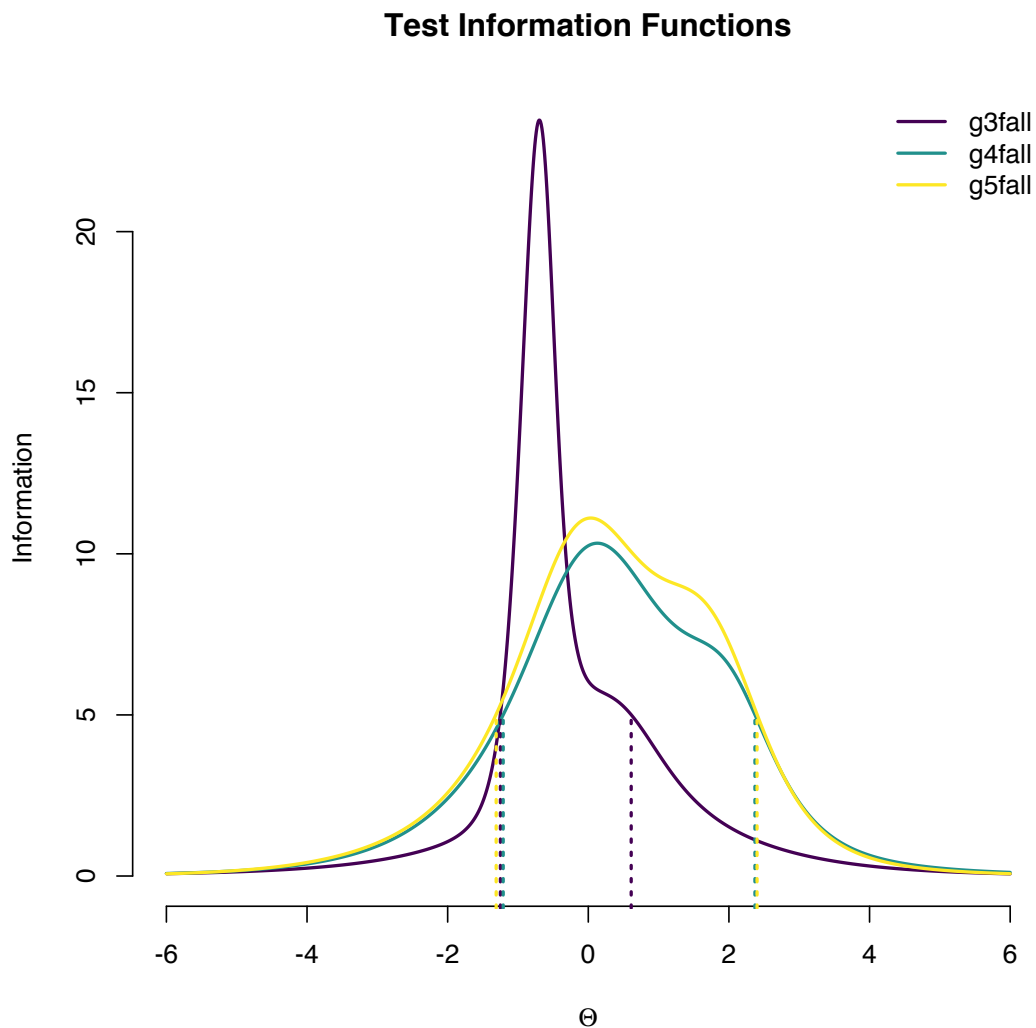


Figure 4.5. Test information functions for grades 3, 4, and 5.

Each test information function above is annotated with two vertical lines, which indicate the lower and upper boundaries of abilities for which the test had a reliability greater than or equal to a Cronbach's alpha level of 0.80.

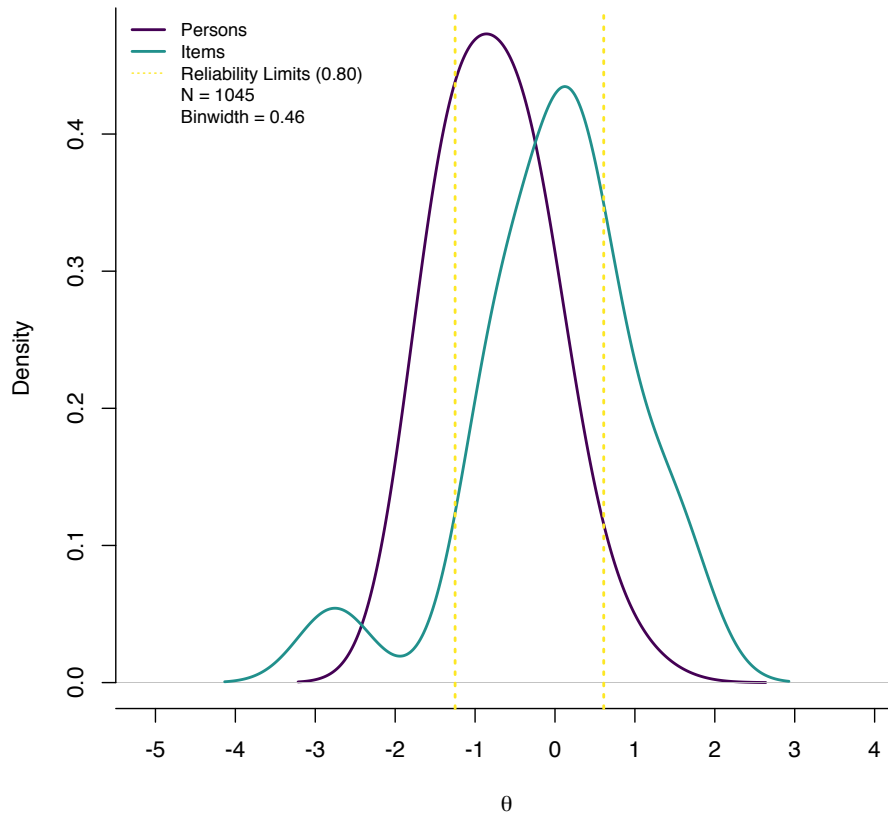


Figure 4.6. Grade 3 item-person plot.

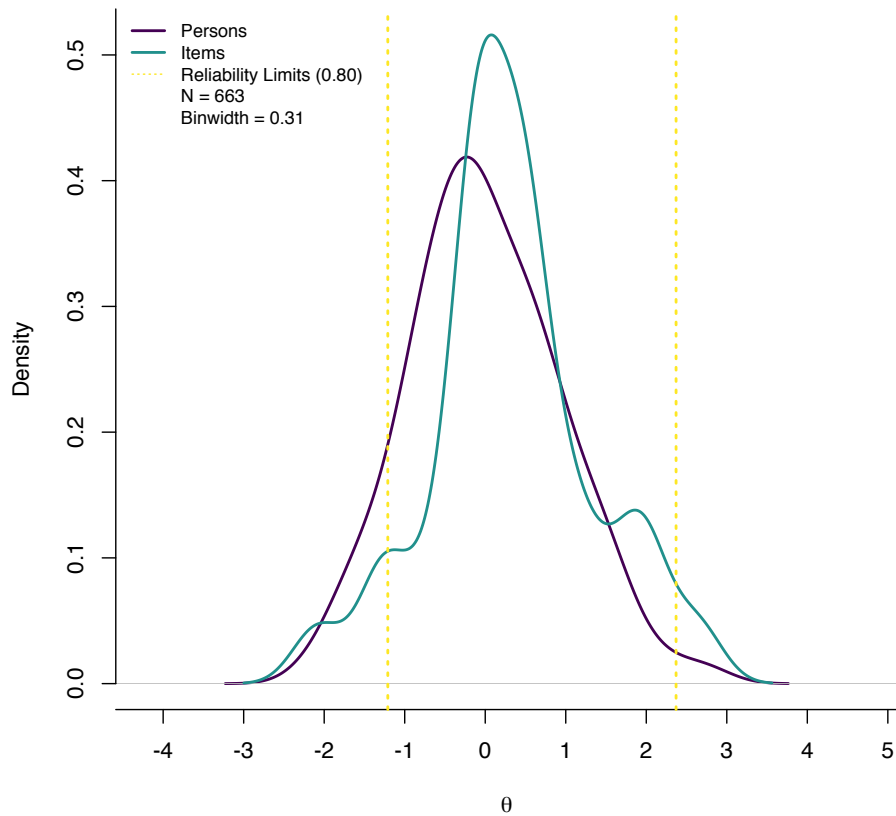


Figure 4.7. Grade 4 item-person plot.

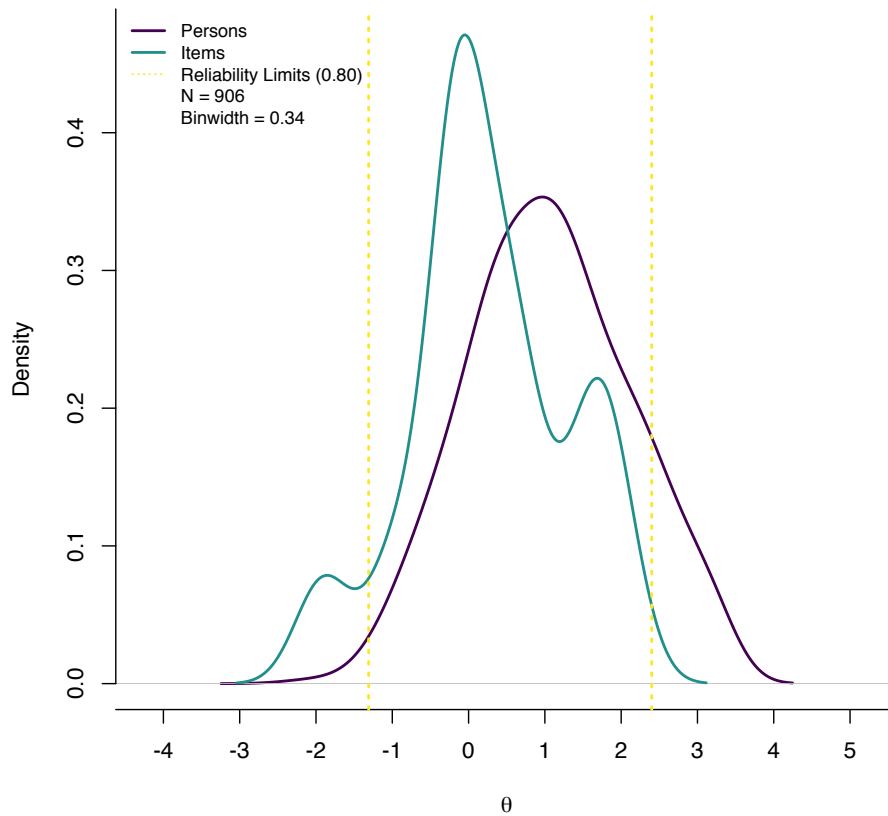


Figure 4.8. Grade 5 item-person plot.

4.4. Predictive Validity

A regression model was used to investigate evidence of predictive validity, with the Fall 2015 3–5 EMSA scores predicting the Spring 2016 3–5 EMSA scores. 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 3–5 EMSA and Spring 2016 3–5 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 3–5 EMSA</i>			
3	843	–0.726	0.647
4	561	0.013	0.938
5	723	1.092	1.063
<i>Spring 2016 3–5 EMSA</i>			
3	843	–0.368	0.915
4	561	–0.009	0.941
5	723	0.748	1.138

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 2,127 grade 3, 4, and 5 students who completed both the fall and spring test, and using SPSS version 24, we found a Pearson correlation of .750 ($p < .001$) between the ability estimates for individual students generated by the Fall 2015 3–5 EMSA test and the ability estimates generated by the Spring 2016 3–5 EMSA for those same students. Therefore, with no adjustment for other factors such as clustering in schools, the student ability estimates from the Fall 2015 K–2 EMSA explain approximately 56.3% of the variance in student scores measured at the end of the school year for these K–2 students.

Table 4.8 shows the fall/spring correlation coefficients and R^2 values disaggregated by grade level. Split by grade level, 843, 561, and 723 students represent grades 3, 4, and 5, respectively. Again using SPSS version 24, we found a Pearson correlation coefficient of .528 for the grade 3 sample, .685 for the grade 4 sample, and .792 for the grade 5 sample. All correlations were statistically significant ($p < .001$). Table 4.9 shows the ICCs calculated for the school and class levels.

Table 4.8. Correlation among Individual Students' Fall 2015 and Spring 2016 EMSA Test Scores

Grade level	Correlation	R ²	Sample <i>n</i>
3	0.528	0.279	843
4	0.685	0.469	561
5	0.792	0.627	723

Note. Models used vertically equated θ estimates for the Fall 2015 EMSA to predict the vertically equated θ estimates for the Spring 2016 estimates. Scores were vertically equated across grade levels within season, but the fall and spring test scores were not equated. All available complete-case data were used. The models did not account for clustering within school or classroom.

Table 4.9. Intraclass Correlation Coefficients, Disaggregated by Grade Level

Grade	ICC
<i>Classroom level</i>	
3	0.034
4	0.125
5	0.186
<i>School level</i>	
3	0.124
4	0.080
5	0.059

Note. ICCs were based on a three-level model, with students at level one, classrooms at level 2, and schools at level 3. The grade 3 sample includes 1,040 students, the grade 4 sample 662, and the grade 5 sample 907.

5. Discussion and Reflection

The Fall 2015 3–5 EMSA tests were the first generation of EMSA tests designed to assess students' mathematical abilities in grades 3–5. In addition, they were the first set of EMSA tests designed to be administered to grade 3–5 students. This task involved 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 overall difficulty of the grade 4 test appeared to align well with the ability levels of the students in the grade 4 sample. The test appears to have adequate internal consistency/reliability across a broad swath of ability levels at grades 4 and 5. Inclusion of a few higher-difficulty items on a future grade 5 test may improve measurement precision in the upper tail of the ability distribution.

The difficulty of the grade 3 test did not align as well with the ability levels of the examinees. The inclusion of a few low-difficulty items on a future version of the grade 3 test may improve alignment and measurement precision in the lower tail of the ability spectrum.

This version of a fall EMSA was the first at the intermediate-grades level. Because we did not know which items would be retained after the field test and initial review, and because we wanted to be sure to have a sufficient number of anchor items on test forms at adjacent grade levels to do the vertical scaling, we made the test forms very similar at the three grade levels, where therefore have similar levels of difficulty. Fifth graders clearly had higher ability levels than fourth graders, and fourth graders than third graders, especially at the higher levels of ability. The test characteristic curves demonstrate the similarity of the difficulty levels of the three tests. Future test forms should include additional high-difficulty items at grade 5 and some low-difficulty items at grade 3. This modification may create more separation in the test characteristic curves and will improve the precision of our estimates of student abilities for students in the top percentiles of student abilities.

The variation in the ICCs for different grade levels is interesting. The classroom-level ICCs were highest for grade 5 students, whereas the school-level ICCs were highest for grade 3 students. The generalizability of this result cannot be determined from this sample. Future samples should be watched for similar results.

Overall, the content review, feasibility study, and results of data analysis support the assertion that the Fall 2015 3–5 EMSA tests are be an adequate assessment tool for their 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.
- Baturo, A. R. (2004). Empowering Andrea to help year 5 students construct fraction understanding. Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education, 2, 95–102.
- Beckmann, S. (2005). *Mathematics for elementary teachers*. Boston, MA: Pearson Education.
- Bright, G. W., Behr, M. J., Post, T. R., & Wachsmuth, I. (1988). Identifying fractions on number lines. *Journal for Research in Mathematics Education, 19*, 215–232.
- Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking Mathematically: Integrating Arithmetic and Algebra in Elementary School*. Portsmouth, NH: Heinemann.
- Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (2015). *Children's mathematics: Cognitively guided instruction* (2nd ed.). Portsmouth, NH: Heinemann.
- Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.
- Dixon, J. K., Larson, M., Leiva, M. A., & Adams, T. L. (2013). *GoMath! Florida*. Orlando, FL: Houghton Mifflin Harcourt.
- Empson, S. & Levi, L. (2011). *Extending Children's Mathematics: Fractions and Decimals*. Portsmouth, NH: Heinemann.
- 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*, 370–378.
- Florida Department of Education (2014). *Mathematics Florida Standards*. Retrieved from <http://www.fldoe.org/core/fileparse.php/5390/urlt/0081015-mathfs.pdf>
- Hackenberg, A., Norton, A., Wilkins, J., & Steffe, L. (2009, April). *Testing hypotheses about students' operational development of fractions*. Paper presented at the Research Pre-session of the National Council of Teachers of Mathematics, Washington, DC.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185. <https://www.doi.org/10.1007/BF02289447>
- Kiearan, T. E. (1976). On the mathematical, cognitive, and instructional foundations of rational numbers. In R. A. Lesh & D. A. Bradbard (Eds.), *Number and measurement: Papers from a research workshop*. Athens, GA: Georgia Center for the Study of Learning and Teaching Mathematics.
- Kolen, M. J., & Brennan, R.L. (2014). *Test equating, scaling, and linking: methods and practices* (3rd ed.). New York: Springer.

- Lamon, S. J. (2005). *Teaching fractions and ratios for understanding: Essential content knowledge and instructional strategies for teachers* (2nded.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Larson, C. N. (1980). Locating proper fractions on number lines: Effect of length and equivalence. *School Science and Mathematics, 80*(5), 423–428.
- Lewis, C. & Perry, R. (2017). Lesson study to scale up research-based knowledge: A randomized, controlled trial of fractions learning. *Journal for Research in Mathematics Education, 48*(3), 261–299. <http://www.jstor.org/stable/10.5951/jresmetheduc.48.3.0261>
- Massachusetts Department of Education (2013). Release of Spring 2013 MCAS Test Items. Retrieved from <http://www.doe.mass.edu/mcas/2013/release/intro.pdf>
- 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.
- Pothier, Y., & Sawada, D. (1983). Partitioning: The emergence of rational number ideas in children. *Journal for Research in Mathematics Education, 14*, 307–317.
- R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>
- Revelle, W. (2017). *psych: Procedures for personality and psychological research* (Version 1.7.5). Evanston, IL: 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
- Saxe, G. B., Diakow, R., & Gearhart, M. (2013). Towards curricular coherence in integers and fractions: A study of the efficacy of a lesson sequence that uses the number line as the principal representational context. *ZDM Mathematics Education, 45*(3), 343–364. <https://www.doi.org/10.1007/s11858-012-0466-2>
- Saxe, G. B., Kirby, K., Kang, B., Le, M., & Schneider, A. (2015). Studying cognition through time in a classroom community: The interplay between “everyday” and “scientific” concepts. *Human Development, 58*, 5–44. <https://www.doi.org/10.1159/000371560>
- 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., Champagne, Z. M., Whitacre, I., & McCrackin, S. (2019). Comparing the frequency and variation of additive word problems in U.S. first-grade textbooks in the 1980s and the Common Core era. Manuscript submitted for publication.
- 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., LaVenia, M., Bauduin, C., & Farina, K. (2016b). Elementary mathematics student assessment: Measuring the performance of grade 1 and 2 students in counting, word problems, and computation in fall 2014 (Research Report No. 2016-04). Tallahassee, FL: Learning Systems Institute, Florida State University. <http://dx.doi.org/10.17125/fsu.1508174887>
- Schoen, R. C., LaVenia, M., Champagne, Z. M., & Farina, K., (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., LaVenia, 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 (Report No. 2016-02). Tallahassee, FL: Learning Systems Institute, Florida State University. <https://doi.org/10.17125/fsu.1493238666>
- Schoen, R. C., Liu, S., Yang, X., & Paek, I. (2017). *Psychometric report for the Early Fractions Test administered with third- and fourth-grade students in fall 2016*. (Research Report No. 2017-10). Tallahassee, FL: Learning Systems Institute, Florida State University. DOI: 10.17125/fsu.1512509662.
- Siegler, R. S., & Lortie-Forgues, H. (2015). Conceptual knowledge of fraction arithmetic. *Journal of Educational Psychology, 107*, 909–918.
- Siegler, R. S., & Pyke, A. A. (2013). Developmental and individual differences in understanding of fractions. *Developmental Psychology, 49*(10), 1994–2004.
- Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*, 273–296.
- 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*, 153–171.
- 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>

Appendix A. Grade 3 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.

Third 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?

- | | | | | |
|-----------------------|-----------------------|----------------------------------|-----------------------|-----------------------|
| 1 | 2 | 3 | 4 | 5 |
| <input type="radio"/> | <input type="radio"/> | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> |

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



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.



1)

[Redacted]

[Redacted]

[Redacted]

2)

[Redacted]

[Redacted]

[Redacted]



3)

[Redacted]

[Redacted]

4)

[Redacted]

[Redacted]

Answer: _____ [Redacted]

5)

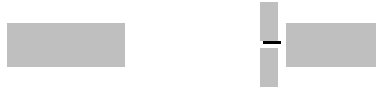
[Redacted]

[Redacted]

[Redacted]



6)



0

0

7)



0

0

0

0

0

8)



0

0

0

0

0



9)

[Redacted]

[Redacted]

[Redacted]

10)

[Redacted]

Answer: _____ [Redacted]



11)



0



0



0



0



0

12)



0



0



0



0



0



13)



0

0



0

0



0

0



0

0



Appendix B. Grade 4 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.

Fourth 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?

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

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



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.



1)

[Redacted]

Answer: _____ [Redacted]

[Redacted]

2)

[Redacted]

Answer: _____ [Redacted]

[Redacted]



3)



4)



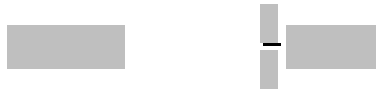
5)



Answer: _____



6)



0

0

7)



0

0

0

0

0

8)



0

0

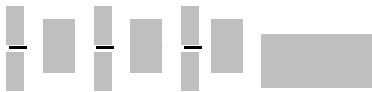
0

0

0



9) For each equation, write the number in the blank that will make the equation correct.



10)

[Redacted]

Answer: _____ [Redacted]

11)

[Redacted]

Answer: _____ [Redacted]



- 12) This dot shows where $\frac{1}{2}$ is on the number line. Draw another dot to show the location of $\frac{1}{4}$.



- 13) There are three dots placed on the following number line. The dots are labeled A, B, and C. Write the number that each dot represents.



A _____

B _____

C _____



14)



15)



Fill in the bubble next to the bar that represents $\frac{3}{5}$ of the whole bar.



16)



Fill in the bubble next to the bar that is most likely to be Pat's bar.



17)



18) For each pair, fill in the bubble under the fraction that is greater. If the two fractions are equivalent, fill in both bubbles.



Appendix C. Grade 5 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.

Fifth 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?

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

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



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.



1)



Answer: _____ mile(s)



2)



Answer: _____ mile(s)



3)

[Redacted]

[Redacted]

[Redacted]

[Redacted]

[Redacted]

[Redacted]

[Redacted]

[Redacted]

[Redacted]

[Redacted] mile



6)



0

0

7)



0

0

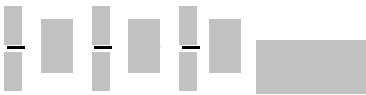
0

0

0



8) For each equation, write the number in the blank that will make the equation correct.



9)



Answer: _____ package(s)

10)



Answer: _____ cup(s)

11)



Answer: _____ mile(s)



12)

[Redacted text]

[Redacted text]

13)

[Redacted text]

[Redacted text]

A _____

B _____

C _____



14)



0



0



0



0



0

15)



0

0

0

0

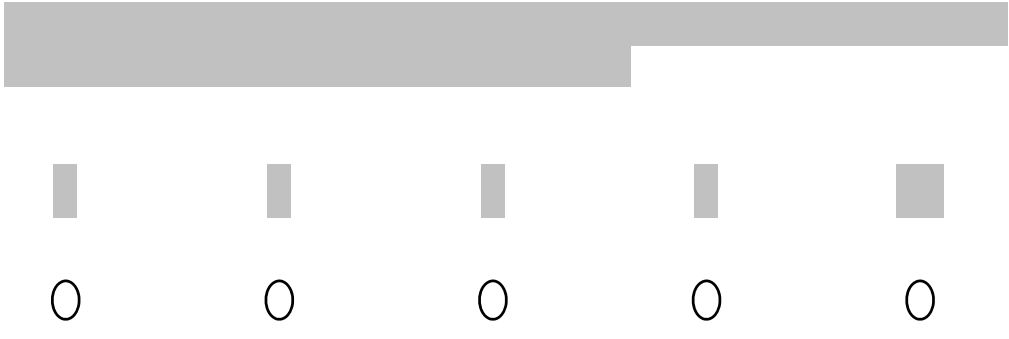
0



16)



17)



18) For each pair, fill in the bubble under the fraction that is greater. If the two fractions are equivalent, fill in both bubbles.



Appendix D. Grade 3 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 Third 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 assessing your mathematics students. This assessment is designed to be group-administered, with students completing the assessment independently. Please administer the Beginning of Year Student Mathematics 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) and administer the assessment as early as possible in the school year.

Items on this assessment are presented in three formats: multiple choice, fill-in-the-blank, and performance tasks. Students should use pencils to bubble, write, and draw their answers. A requested script for the test administrator to use during administration begins on page 5 of this guide. The script should be followed 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 Year Student Assessment Materials section on page 4.

Student Assessment Window

Student assessment should occur according to the following schedule:

School District	Testing Window
Bay District Schools	August 18, 2015 – August 28, 2015
Broward County School District	August 24, 2015 – September 4, 2015
FSU Lab School	August 17, 2015 – August 28, 2015
Holmes County School District	August 12, 2015 – August 26, 2015
Jackson County School District	August 10, 2015 – August 24, 2015
Leon County School District	August 17, 2015 – August 28, 2015
Okaloosa County School District	August 17, 2015 – August 28, 2015
Orange County School District	August 24, 2015 – September 4, 2015
Seminole County School District	August 17, 2015 – August 28, 2015
Sumter County School District	August 10, 2015 – August 24, 2015
Taylor County School District	August 10, 2015 – August 24, 2015
Wakulla County School District	August 20, 2015 – September 3, 2015
Walton County School District	August 10, 2015 – August 24, 2015

Materials

The following materials are required for testing:

- Administration Instructions for the Third Grade Beginning of Year Student Mathematics Assessment (this document)
- A test booklet for each student (one per student, provided)
- At least one sharpened pencil for each student

Test Booklets

The students should bubble, write, and draw their answers directly in the test booklets. Should you need additional testing materials, please contact Amanda Tazaz (atazaz@lsi.fsu.edu). Remember that these materials are to remain at the school site until the testing window has ended. The

materials should be stored in a secure, access-restricted location at all times.

Students to be Tested

We ask that you administer the assessment to students for whom you are the teacher of record. Therefore, if you teach mathematics to multiple groups of students, 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, school, teacher, 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 Year Student Assessment for the *Foundations for Success in STEM* Study is designed to be group-administered, with students completing the assessment independently. 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 Year Student Assessment guidelines and administration instructions.
- No talking or communication between students is permitted during testing.
- Students are permitted to use mathematics manipulatives during the assessment if they would ordinarily be permitted to use manipulatives in your classroom.
- If individual students have difficulty with reading items, it is permissible to read the questions to the students. If you read the items for the student(s), avoid emphasizing words in ways that give extra clues about what to pay attention to in the items.
- Avoid answering student questions in ways that offer clues about how to approach problems. Student responses should reflect their current math knowledge.

Administering the Beginning of 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 Year Student Assessment should be consistent with the testing conditions for other student assessments administered in the classroom. For example, desks should be spaced or student “privacy folders” used if that is what you would usually do.

To ensure that the students’ test responses are valid, it is important that appropriate procedures are

followed when administering the Beginning of Year Student Assessment. These procedures include:

- Administration of the appropriate test level (Grade 3 assessment for Grade 3 students)
- Adherence to the Beginning of 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 testing accommodations are specified in their plans.

Testing Time Allocation

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

Submitting the Beginning of Year Student Assessment Materials

Upon conclusion of testing, repack the test booklets (used and unused) in the original packaging. Also, please be sure to include this document (Administration Instructions for the Third Grade Beginning of Year Student Mathematics Assessment) and your completed student information sheet (located at end of this packet). A member of the project staff will coordinate with your school to set a date to retrieve the testing materials.

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):

School District	Target Pick-up Window
Bay District Schools	August 31, 2015 – September 4, 2015
Broward County School District	September 7, 2015 – September 11, 2015
FSU Lab School	August 31, 2015 – September 4, 2015
Holmes County School District	August 27, 2015 – September 3, 2015
Jackson County School District	August 24, 2015 – August 28, 2015
Leon County School District	August 31, 2015 – September 4, 2015
Okaloosa County School District	August 31, 2015 – September 4, 2015
Orange County School District	September 7, 2015 – September 11, 2015
Seminole County School District	August 31, 2015 – September 4, 2015
Sumter County School District	August 24, 2015 – August 28, 2015
Taylor County School District	August 24, 2015 – August 28, 2015
Wakulla County School District	September 7, 2015 – September 11, 2015
Walton County School District	August 24, 2015 – August 28, 2015

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

Assessment Administration Instructions – Grade 3

[The boxes contain the script that you will read to the students at the time you administer the assessment.]

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.

Distribute the assessments to students, ensuring that each student receives the test booklet that has been pre-labeled with his or her name.

It is your choice if you want to answer the questions on this assessment. Some kinds of problems may be new or challenging for you. Don't worry if you are not sure how to solve them. You won't be graded on this assessment. Just try your best.

This assessment will ask you to give your answers in some different ways. Sometimes you will be asked to write your answer on a blank line. Other times you will need to give your answer by marking a spot on a number line or shading part of a shape. Some questions include multiple answer choices. For these, you will need to fill in the bubble under the choice you think is the correct answer. The first page of the assessment provides a sample of how you will mark your answers for multiple choice questions.

Look at the sample question on the front of your booklet.

It asks: 'What grade are you in?' The correct answer choice is 3, for Third Grade. Notice how the bubble under the 3 has been filled in for you. You are going to mark your answer choices for multiple choice questions the same way, by filling in the bubble below the answer choice you think is correct.

For each question, I would like for you to try hard to figure out the answer. If you are not sure, it is okay to make a guess.

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.

When you have completed the assessment, you may check back over your answers. I will collect the assessments when everyone is done.

Are there any questions?

Address any questions.

If there are no more questions, you may open the assessment and begin.

Circulate, ensuring that all students are attending to the items presented on the front and back of pages in the assessment.

As a reminder, if individual students are having difficulty reading items or parts of items, it is okay to help with reading. But do not read in ways that give clues about what to pay attention to in the items, and do not help with the math.

As students finish, encourage them to check their work and check to make sure that no items were overlooked (particularly those on the back of pages).

We anticipate that the assessment will take students approximately 45 minutes. Students should generally be allowed to take the time they need to finish the assessment. Use your discretion to determine when students have stopped working productively. When all students have finished or you have determined the testing period is over, you should then collect all of the student assessments.

Beginning of the Year Student Information Sheet

INSTRUCTION: Please enter the information at the top of this form and provide the following information for all students in your class. For each student, provide his or her unique district ID #, first and last name as it appears on official records, indication of whether a completed assessment is enclosed, and any other relevant notes. Notes are optional; all other information is required.

School Name:		Testing Date:	
Teacher Name:		Testing Start Time:	
Grade Level(s):		Testing End Time:	
Were mathematics manipulatives used by students during the assessment? (circle one)			YES or NO

Student's District ID	Student's First Name	Student's Last Name	Student's Nickname (if any)	Completed Assessment Enclosed (circle one)	ELL or Testing Accommodations?	Notes
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		

Student's District ID	Student's First Name	Student's Last Name	Student's Nickname (if any)	Completed Assessment Enclosed (circle one)	ELL or Testing Accommodations?	Notes
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		

Appendix E. Grade 4 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 Fourth 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 assessing your mathematics students. This assessment is designed to be group-administered, with students completing the assessment independently. Please administer the Beginning of Year Student Mathematics 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) and administer the assessment as early as possible in the school year.

Items on this assessment are presented in three formats: multiple choice, fill-in-the-blank, and performance tasks. Students should use pencils to bubble, write, and draw their answers. A requested script for the test administrator to use during administration begins on page 5 of this guide. The script should be followed 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 Year Student Assessment Materials section on page 4.

Student Assessment Window

Student assessment should occur according to the following schedule:

School District	Testing Window
Bay District Schools	August 18, 2015 – August 28, 2015
Broward County School District	August 24, 2015 – September 4, 2015
FSU Lab School	August 17, 2015 – August 28, 2015
Holmes County School District	August 12, 2015 – August 26, 2015
Jackson County School District	August 10, 2015 – August 24, 2015
Leon County School District	August 17, 2015 – August 28, 2015
Okaloosa County School District	August 17, 2015 – August 28, 2015
Orange County School District	August 24, 2015 – September 4, 2015
Seminole County School District	August 17, 2015 – August 28, 2015
Sumter County School District	August 10, 2015 – August 24, 2015
Taylor County School District	August 10, 2015 – August 24, 2015
Wakulla County School District	August 20, 2015 – September 3, 2015
Walton County School District	August 10, 2015 – August 24, 2015

Materials

The following materials are required for testing:

- Administration Instructions for the Fourth Grade Beginning of Year Student Mathematics Assessment (this document)
- A test booklet for each student (one per student, provided)
- At least one sharpened pencil for each student

Test Booklets

The students should bubble, write, and draw their answers directly in the test booklets. Should you need additional testing materials, please contact Amanda Tazaz (atazaz@lsi.fsu.edu). Remember that these materials are to remain at the school site until the testing window has ended. The

materials should be stored in a secure, access-restricted location at all times.

Students to be Tested

We ask that you administer the assessment to students for whom you are the teacher of record. Therefore, if you teach mathematics to multiple groups of students, 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, school, teacher, 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 Year Student Assessment for the *Foundations for Success in STEM* Study is designed to be group-administered, with students completing the assessment independently. 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 Year Student Assessment guidelines and administration instructions.
- No talking or communication between students is permitted during testing.
- Students are permitted to use mathematics manipulatives during the assessment if they would ordinarily be permitted to use manipulatives in your classroom.
- If individual students have difficulty with reading items, it is permissible to read the questions to the students. If you read the items for the student(s), avoid emphasizing words in ways that give extra clues about what to pay attention to in the items.
- Avoid answering student questions in ways that offer clues about how to approach problems. Student responses should reflect their current math knowledge.

Administering the Beginning of 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 Year Student Assessment should be consistent with the testing conditions for other student assessments administered in the classroom. For example, desks should be spaced or student “privacy folders” used if that is what you would usually do.

To ensure that the students’ test responses are valid, it is important that appropriate procedures are

followed when administering the Beginning of Year Student Assessment. These procedures include:

- Administration of the appropriate test level (Grade 4 assessment for Grade 4 students)
- Adherence to the Beginning of 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 testing accommodations are specified in their plans.

Testing Time Allocation

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

Submitting the Beginning of Year Student Assessment Materials

Upon conclusion of testing, repack the test booklets (used and unused) in the original packaging. Also, please be sure to include this document (Administration Instructions for the Fourth Grade Beginning of Year Student Mathematics Assessment) and your completed student information sheet (located at end of this packet). A member of the project staff will coordinate with your school to set a date to retrieve the testing materials.

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):

School District	Target Pick-up Window
Bay District Schools	August 31, 2015 – September 4, 2015
Broward County School District	September 7, 2015 – September 11, 2015
FSU Lab School	August 31, 2015 – September 4, 2015
Holmes County School District	August 27, 2015 – September 3, 2015
Jackson County School District	August 24, 2015 – August 28, 2015
Leon County School District	August 31, 2015 – September 4, 2015
Okaloosa County School District	August 31, 2015 – September 4, 2015
Orange County School District	September 7, 2015 – September 11, 2015
Seminole County School District	August 31, 2015 – September 4, 2015
Sumter County School District	August 24, 2015 – August 28, 2015
Taylor County School District	August 24, 2015 – August 28, 2015
Wakulla County School District	September 7, 2015 – September 11, 2015
Walton County School District	August 24, 2015 – August 28, 2015

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

Assessment Administration Instructions – Grade 4

[The boxes contain the script that you will read to the students at the time you administer the assessment.]

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.

Distribute the assessments to students, ensuring that each student receives the test booklet that has been pre-labeled with his or her name.

It is your choice if you want to answer the questions on this assessment. Some kinds of problems may be new or challenging for you. Don't worry if you are not sure how to solve them. You won't be graded on this assessment. Just try your best.

This assessment will ask you to give your answers in some different ways. Sometimes you will be asked to write your answer on a blank line. Other times you will need to give your answer by marking a spot on a number line or shading part of a shape. Some questions include multiple answer choices. For these, you will need to fill in the bubble under the choice you think is the correct answer. The first page of the assessment provides a sample of how you will mark your answers for multiple choice questions.

Look at the sample question on the front of your booklet.

It asks: 'What grade are you in?' The correct answer choice is 4, for Fourth Grade. Notice how the bubble under the 4 has been filled in for you. You are going to mark your answer choices for multiple choice questions the same way, by filling in the bubble below the answer choice you think is correct.

For each question, I would like for you to try hard to figure out the answer. If you are not sure, it is okay to make a guess.

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.

When you have completed the assessment, you may check back over your answers. I will collect the assessments when everyone is done.

Are there any questions?

Address any questions.

If there are no more questions, you may open the assessment and begin.

Circulate, ensuring that all students are attending to the items presented on the front and back of pages in the assessment.

As a reminder, if individual students are having difficulty reading items or parts of items, it is okay to help with reading. But do not read in ways that give clues about what to pay attention to in the items, and do not help with the math.

As students finish, encourage them to check their work and check to make sure that no items were overlooked (particularly those on the back of pages).

We anticipate that the assessment will take students approximately 45 minutes. Students should generally be allowed to take the time they need to finish the assessment. Use your discretion to determine when students have stopped working productively. When all students have finished or you have determined the testing period is over, you should then collect all of the student assessments.

Beginning of the Year Student Information Sheet

INSTRUCTION: Please enter the information at the top of this form and provide the following information for all students in your class. For each student, provide his or her unique district ID #, first and last name as it appears on official records, indication of whether a completed assessment is enclosed, and any other relevant notes. Notes are optional; all other information is required.

School Name:		Testing Date:	
Teacher Name:		Testing Start Time:	
Grade Level(s):		Testing End Time:	
Were mathematics manipulatives used by students during the assessment? (circle one)			YES or NO

Student's District ID	Student's First Name	Student's Last Name	Student's Nickname (if any)	Completed Assessment Enclosed (circle one)	ELL or Testing Accommodations?	Notes
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		

Student's District ID	Student's First Name	Student's Last Name	Student's Nickname (if any)	Completed Assessment Enclosed (circle one)	ELL or Testing Accommodations?	Notes
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		

Appendix F. Grade 5 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 Fifth 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 assessing your mathematics students. This assessment is designed to be group-administered, with students completing the assessment independently. Please administer the Beginning of Year Student Mathematics 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) and administer the assessment as early as possible in the school year.

Items on this assessment are presented in three formats: multiple choice, fill-in-the-blank, and performance tasks. Students should use pencils to bubble, write, and draw their answers. A requested script for the test administrator to use during administration begins on page 5 of this guide. The script should be followed 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 Year Student Assessment Materials section on page 4.

Student Assessment Window

Student assessment should occur according to the following schedule:

School District	Testing Window
Bay District Schools	August 18, 2015 – August 28, 2015
Broward County School District	August 24, 2015 – September 4, 2015
FSU Lab School	August 17, 2015 – August 28, 2015
Holmes County School District	August 12, 2015 – August 26, 2015
Jackson County School District	August 10, 2015 – August 24, 2015
Leon County School District	August 17, 2015 – August 28, 2015
Okaloosa County School District	August 17, 2015 – August 28, 2015
Orange County School District	August 24, 2015 – September 4, 2015
Seminole County School District	August 17, 2015 – August 28, 2015
Sumter County School District	August 10, 2015 – August 24, 2015
Taylor County School District	August 10, 2015 – August 24, 2015
Wakulla County School District	August 20, 2015 – September 3, 2015
Walton County School District	August 10, 2015 – August 24, 2015

Materials

The following materials are required for testing:

- Administration Instructions for the Fifth Grade Beginning of Year Student Mathematics Assessment (this document)
- A test booklet for each student (one per student, provided)
- At least one sharpened pencil for each student

Test Booklets

The students should bubble, write, and draw their answers directly in the test booklets. Should you need additional testing materials, please contact Amanda Tazaz (atazaz@lsi.fsu.edu). Remember

that these materials are to remain at the school site until the testing window has ended. The materials should be stored in a secure, access-restricted location at all times.

Students to be Tested

We ask that you administer the assessment to students for whom you are the teacher of record. Therefore, if you teach mathematics to multiple groups of students, 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, school, teacher, 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 Year Student Assessment for the *Foundations for Success in STEM* Study is designed to be group-administered, with students completing the assessment independently. 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 Year Student Assessment guidelines and administration instructions.
- No talking or communication between students is permitted during testing.
- Students are permitted to use mathematics manipulatives during the assessment if they would ordinarily be permitted to use manipulatives in your classroom.
- If individual students have difficulty with reading items, it is permissible to read the questions to the students. If you read the items for the student(s), avoid emphasizing words in ways that give extra clues about what to pay attention to in the items.
- Avoid answering student questions in ways that offer clues about how to approach problems. Student responses should reflect their current math knowledge.

Administering the Beginning of 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 Year Student Assessment should be consistent with the testing conditions for other student assessments administered in the classroom. For example, desks should be spaced or student “privacy folders” used if that is what you would usually do.

To ensure that the students’ test responses are valid, it is important that appropriate procedures are

followed when administering the Beginning of Year Student Assessment. These procedures include:

- Administration of the appropriate test level (Grade 5 assessment for Grade 5 students)
- Adherence to the Beginning of 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 testing accommodations are specified in their plans.

Testing Time Allocation

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

Submitting the Beginning of Year Student Assessment Materials

Upon conclusion of testing, repack the test booklets (used and unused) in the original packaging. Also, please be sure to include this document (Administration Instructions for the Fifth Grade Beginning of Year Student Mathematics Assessment) and your completed student information sheet (located at end of this packet). A member of the project staff will coordinate with your school to set a date to retrieve the testing materials.

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):

School District	Target Pick-up Window
Bay District Schools	August 31, 2015 – September 4, 2015
Broward County School District	September 7, 2015 – September 11, 2015
FSU Lab School	August 31, 2015 – September 4, 2015
Holmes County School District	August 27, 2015 – September 3, 2015
Jackson County School District	August 24, 2015 – August 28, 2015
Leon County School District	August 31, 2015 – September 4, 2015
Okaloosa County School District	August 31, 2015 – September 4, 2015
Orange County School District	September 7, 2015 – September 11, 2015
Seminole County School District	August 31, 2015 – September 4, 2015
Sumter County School District	August 24, 2015 – August 28, 2015
Taylor County School District	August 24, 2015 – August 28, 2015
Wakulla County School District	September 7, 2015 – September 11, 2015
Walton County School District	August 24, 2015 – August 28, 2015

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

Assessment Administration Instructions – Grade 5

[The boxes contain the script that you will read to the students at the time you administer the assessment.]

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.

Distribute the assessments to students, ensuring that each student receives the test booklet that has been pre-labeled with his or her name.

It is your choice if you want to answer the questions on this assessment. Some kinds of problems may be new or challenging for you. Don't worry if you are not sure how to solve them. You won't be graded on this assessment. Just try your best.

This assessment will ask you to give your answers in some different ways. Sometimes you will be asked to write your answer on a blank line. Other times you will need to give your answer by marking a spot on a number line or shading part of a shape. Some questions include multiple answer choices. For these, you will need to fill in the bubble under the choice you think is the correct answer. The first page of the assessment provides a sample of how you will mark your answers for multiple choice questions.

Look at the sample question on the front of your booklet.

It asks: 'What grade are you in?' The correct answer choice is 5, for Fifth Grade. Notice how the bubble under the 5 has been filled in for you. You are going to mark your answer choices for multiple choice questions the same way, by filling in the bubble below the answer choice you think is correct.

For each question, I would like for you to try hard to figure out the answer. If you are not sure, it is okay to make a guess.

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.

When you have completed the assessment, you may check back over your answers. I will collect the assessments when everyone is done.

Are there any questions?

Address any questions.

If there are no more questions, you may open the assessment and begin.

Circulate, ensuring that all students are attending to the items presented on the front and back of pages in the assessment.

As a reminder, if individual students are having difficulty reading items or parts of items, it is okay to help with reading. But do not read in ways that give clues about what to pay attention to in the items, and do not help with the math.

As students finish, encourage them to check their work and check to make sure that no items were overlooked (particularly those on the back of pages).

We anticipate that the assessment will take students approximately 45 minutes. Students should generally be allowed to take the time they need to finish the assessment. Use your discretion to determine when students have stopped working productively. When all students have finished or you have determined the testing period is over, you should then collect all of the student assessments.

Beginning of the Year Student Information Sheet

INSTRUCTION: Please enter the information at the top of this form and provide the following information for all students in your class. For each student, provide his or her unique district ID #, first and last name as it appears on official records, indication of whether a completed assessment is enclosed, and any other relevant notes. Notes are optional; all other information is required.

School Name:		Testing Date:	
Teacher Name:		Testing Start Time:	
Grade Level(s):		Testing End Time:	
Were mathematics manipulatives used by students during the assessment? (circle one)			YES or NO

Student's District ID	Student's First Name	Student's Last Name	Student's Nickname (if any)	Completed Assessment Enclosed (circle one)	ELL or Testing Accommodations?	Notes
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		

Student's District ID	Student's First Name	Student's Last Name	Student's Nickname (if any)	Completed Assessment Enclosed (circle one)	ELL or Testing Accommodations?	Notes
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		
				YES or NO		

Appendix G. Scoring Key

Items were scored by means of both the scoring key shown in Table G.1 and an overlay that follows.

Table G.1. Grade 3 Scoring Key

Item	Item description	Response format	Data entry	Correct response
G3G4G5i1a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4G5i1b	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i2a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4G5i2b	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i3	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i4	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3i5_G4G5i12	[REDACTED]	Constructed	This overlay is divided into segments. Enter the letter that corresponds to the segment the student’s mark resides within. Enter DNS for did not solve or UI for unclear intent.	Scored using overlay
G3G4G5i6	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3G4G5i7	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3i8_G4G5i14	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3G4i9a_G5i8a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4i9b_G5i8b	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4i10_G5i9	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3i11_G4i8	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3i12_G4G5i17	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3i13a_G4G5i18a	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3i13b_G4G5i18b	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3i13c_G4G5i18c	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3i13d_G4G5i18d	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]

Table G.2. Grade 4 Scoring Key

Item	Item description	Response format	Data entry	Correct response
G3G4G5i1a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4G5i1b	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i2a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4G5i2b	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i3	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i4	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4G5i5	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4G5i6	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3G4G5i7	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3i11_G4i8	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3G4i9a_G5i8a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4i9b_G5i8b	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4i9c_G5i8c	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4i9d_G5i8d	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4i9e_G5i8e	[REDACTED] = □	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4i10_G5i9	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4i11_G5i10	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3i5_G4G5i12	[REDACTED]	Constructed	This overlay is divided into segments. Enter the letter that corresponds to the segment the student’s mark resides within. Enter DNS for did not solve or UI for unclear intent.	Scored using overlay
G4G5i13a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4G5i13b	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4G5i13c	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]

G3i8_G4G5i14	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G4G5i15	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	D position
G4G5i16	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	B position
G3i12_G4G5i17	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G3i13a_G4G5i18a	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G3i13b_G4G5i18b	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G3i13c_G4G5i18c	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G3i13d_G4G5i18d	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	

Table G.3. Grade 5 Scoring Key

Item	Item description	Response format	Data entry	Correct response
G3G4G5i1a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4G5i1b	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i2a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4G5i2b	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i3	[REDACTED]	Constructed	Enter a 0 for incorrect, 1 for correct, or DNS for did not solve	Scored using overlay
G3G4G5i4	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4G5i5	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4G5i6	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3G4G5i7	[REDACTED]	Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	[REDACTED]
G3i11_G4i8	[REDACTED]	Selected	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4i9a_G5i8a	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G3G4i9b_G5i8b	1 [REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]
G4i9c_G5i8c	[REDACTED]	Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	[REDACTED]

G4i9d_G5i8d		Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	
G4i9e_G5i8e		Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	
G3G4i10_G5i9		Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	
G4i11_G5i10		Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	
G3i5_G4G5i12		Constructed	This overlay is divided into segments. Enter the letter that corresponds to the segment the student’s mark resides within. Enter DNS for did not solve or UI for unclear intent.	Scored using overlay
G4G5i13a		Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	
G4G5i13b		Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	
G4G5i13c		Constructed	Enter exactly as written. Enter DNS for did not solve or UI for unclear intent.	
G3i8_G4G5i14		Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G4G5i15		Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	D position
G4G5i16		Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	B position
G3i12_G4G5i17		Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G3i13a_G4G5i18a		Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G3i13b_G4G5i18b		Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G3i13c_G4G5i18c		Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	
G3i13d_G4G5i18d		Selected	Enter the number that corresponds to the position of the student’s selected response. Enter DNS for did not solve, UI for unclear intent, or MR for multiple responses.	



Appendix H. Results of Initial Screening

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

H.1. Item-level Statistics

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

Table H.1. Item Statistics for the Grade 3 Test Based on the Grade 3 Sample ($n = 1,045$)

Item	Item description	CTT-based statistics		IRT-based statistics			
		PC (se)	PB	Vertical scale		Within-grade-level scale	
				Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
G3G4G5i1a		.53 (.015)	.53	5.77 (.524)	-0.82 (.178)	4.21 (.524)	-0.40 (.178)
G3G4G5i1b		.44 (.015)	.57	7.20 (.843)	-0.66 (.226)	5.24 (.843)	0.64 (.226)
G3G4G5i2a		.11 (.010)	.49	2.57 (.204)	0.49 (.224)	1.87 (.204)	3.19 (.224)
G3G4G5i2b		.24 (.013)	.54	2.54 (.170)	-0.08 (.132)	1.85 (.170)	1.72 (.132)
G3G4G5i3		.48 (.015)	.42	0.94 (.085)	-0.65 (.068)	0.68 (.085)	0.10 (.068)
G3G4G5i4		.15 (.011)	.47	1.41 (.120)	0.69 (.117)	1.02 (.120)	2.02 (.117)
G3i5_G4G5i12		.22 (.013)	.35	0.90 (.094)	0.80 (.086)	0.65 (.094)	1.39 (.086)
G3G4G5i6		.51 (.015)	.39	0.82 (.082)	-0.83 (.067)	0.60 (.082)	-0.06 (.067)
G3G4G5i7		.12 (.010)	.40	0.94 (.113)	1.55 (.113)	0.68 (.113)	2.16 (.113)
G3i8_G4G5i14		.28 (.014)	.43	0.95 (.090)	0.34 (.078)	0.69 (.090)	1.04 (.078)
G3G4i9a_G5i8a		.29 (.014)	.45	1.09 (.094)	0.15 (.079)	0.80 (.094)	0.98 (.079)
G3G4i9b_G5i8b		.31 (.014)	.49	1.22 (.098)	-0.01 (.080)	0.89 (.098)	0.90 (.080)
G3G4i10_G5i9		.29 (.014)	.47	1.32 (.101)	0.01 (.084)	0.96 (.101)	1.01 (.084)
G3i11_G4i8		.22 (.013)	.33	0.61 (.088)	1.45 (.080)	0.45 (.088)	1.35 (.080)
G3i12_G4G5i17		.38 (.015)	.38	0.60 (.078)	0.12 (.070)	0.44 (.078)	0.52 (.070)
G3i13a_G4G5i18a		.74 (.014)	.21	0.53 (.086)	-2.75 (.086)	0.38 (.086)	-1.06 (.086)
<i>G3i13b_G4G5i18b</i>		<i>.19 (.012)</i>	<i>.24</i>	—	—	—	—
<i>G3i13c_G4G5i18c</i>		<i>.17 (.012)</i>	<i>.09</i>	—	—	—	—
<i>G3i13d_G4G5i18d</i>		<i>.17 (.012)</i>	<i>.28</i>	—	—	—	—

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 4 Test Based on the Grade 4 Sample (n = 663)

Item	Item description	CTT-based statistics		IRT-based statistics			
		PC (se)	PB	Vertical scale		Within-grade-level scale	
				Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
G3G4G5i1a		.67 (.018)	.49	1.30 (.148)	-0.74 (.116)	1.30 (.148)	-0.97 (.116)
G3G4G5i1b		.59 (.019)	.48	1.16 (.132)	-0.42 (.102)	1.16 (.132)	-0.49 (.102)
G3G4G5i2a		.37 (.019)	.61	1.85 (.182)	0.45 (.130)	1.85 (.182)	0.82 (.130)
G3G4G5i2b		.39 (.019)	.49	1.19 (.130)	0.49 (.103)	1.19 (.130)	0.58 (.103)
G3G4G5i3		.77 (.016)	.41	1.24 (.154)	-1.27 (.134)	1.24 (.154)	-1.57 (.134)
G3G4G5i4		.55 (.019)	.60	1.98 (.201)	-0.19 (.129)	1.98 (.201)	-0.38 (.129)
G4G5i5		.51 (.019)	.59	1.79 (.179)	-0.07 (.120)	1.79 (.179)	-0.12 (.120)
G3G4G5i6		.71 (.018)	.35	0.82 (.117)	-1.26 (.102)	0.82 (.117)	-1.03 (.102)
G3G4G5i7		.48 (.019)	.47	2.24 (.224)	0.05 (.136)	2.24 (.224)	0.11 (.136)
G3i11_G4i8		.41 (.019)	.38	0.75 (.103)	0.55 (.089)	0.75 (.103)	0.41 (.089)
G3G4i9a_G5i8a		.50 (.019)	.41	0.82 (.107)	-0.03 (.089)	0.82 (.107)	-0.02 (.089)
G3G4i9b_G5i8b		.51 (.019)	.52	1.28 (.135)	-0.03 (.102)	1.28 (.135)	-0.04 (.102)
G4i9c_G5i8c		.29 (.018)	.59	1.83 (.185)	0.76 (.147)	1.83 (.185)	1.39 (.147)
G4i9d_G5i8d		.39 (.019)	.49	1.20 (.131)	0.48 (.104)	1.20 (.131)	0.57 (.104)
G4i9e_G5i8e		.05 (.009)	.41	2.41 (.386)	2.05 (4.94)	2.41 (.386)	0.56 (4.94)
G3G4i10_G5i9		.41 (.019)	.47	1.02 (.118)	0.43 (.097)	1.02 (.118)	0.44 (.097)
G4i11_G5i10		.29 (.018)	.59	0.93 (.117)	1.14 (.105)	0.93 (.117)	1.06 (.105)
G3i5_G4G5i12		.28 (.017)	.38	0.83 (.113)	1.32 (.103)	0.83 (.113)	1.10 (.103)
G4G5i13a		.12 (.013)	.44	1.59 (.101)	1.77 (.091)	0.46 (.101)	-0.95 (.091)
G4G5i13b		.52 (.019)	.40	0.82 (.173)	-0.14 (.115)	1.66 (.173)	-0.14 (.115)
G4G5i13c		.06 (.009)	.43	2.52 (.121)	1.93 (.096)	1.01 (.121)	0.42 (.096)
G3i8_G4G5i14		.53 (.019)	.56	1.55 (.157)	-0.13 (.112)	1.55 (.157)	-0.21 (.112)
G4G5i15		.20 (.015)	.27	0.57 (.112)	2.64 (.108)	0.57 (.112)	1.50 (.108)
G4G5i16		.28 (.017)	.52	1.41 (.144)	0.90 (.126)	1.41 (.144)	1.27 (.126)
G3i12_G4G5i17		.44 (.019)	.37	0.70 (.101)	0.38 (.087)	0.70 (.101)	0.27 (.087)
G3i13a_G4G5i18a		.71 (.018)	.26	0.46 (.101)	-2.08 (.091)	0.46 (.101)	-0.95 (.091)
G3i13b_G4G5i18b		.52 (.019)	.57	1.66 (.173)	-0.09 (.115)	1.66 (.173)	-0.14 (.115)
G3i13c_G4G5i18c		.41 (.019)	.43	1.01 (.121)	0.41 (.096)	1.01 (.121)	0.42 (.096)
G3i13d_G4G5i18d		.51 (.019)	.55	1.57 (.165)	-0.07 (.112)	1.57 (.165)	-0.10 (.112)

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 5 Test Based on the Grade 5 Sample (n = 906)

Item	Item description	CTT-based statistics		IRT-based statistics			
		PC (se)	PB	Vertical scale		Within-grade-level scale	
				Discrim (se)	Diff (se)	Discrim (se)	Diff (se)
G3G4G5i1a		.80 (.013)	.51	1.46 (.157)	-0.34 (.146)	1.62 (.157)	-2.03 (.146)
G3G4G5i1b		.71 (.015)	.51	1.18 (.123)	0.50 (.104)	1.31 (.123)	-1.18 (.104)
G3G4G5i2a		.63 (.016)	.62	1.71 (.156)	0.52 (.116)	1.89 (.156)	-0.91 (.116)
G3G4G5i2b		.56 (.016)	.51	1.09 (.108)	0.74 (.087)	1.21 (.108)	-0.33 (.087)
G3G4G5i3		.92 (.009)	.30	1.02 (.165)	-1.78 (.182)	1.14 (.165)	-2.90 (.182)
G3G4G5i4		.82 (.013)	.56	1.96 (.206)	-0.23 (.194)	2.17 (.206)	-2.51 (.194)
G4G5i5		.78 (.014)	.51	1.44 (.149)	-0.19 (.133)	1.59 (.149)	-1.78 (.133)
G3G4G5i6		.90 (.010)	.28	0.81 (.140)	-2.01 (.146)	0.90 (.140)	-2.49 (.146)
G3G4G5i7		.82 (.013)	.59	2.46 (.271)	-0.18 (.257)	2.73 (.271)	-3.04 (.257)
G3G4i9a_G5i8a		.74 (.015)	.48	1.15 (.123)	-0.14 (.108)	1.28 (.123)	-1.38 (.108)
G3G4i9b_G5i8b		.74 (.015)	.53	1.40 (.141)	-0.03 (.122)	1.55 (.141)	-1.51 (.122)
G4i9c_G5i8c		.85 (.012)	.50	1.69 (.189)	-0.51 (.190)	1.87 (.189)	-2.63 (.190)
G4i9d_G5i8d		.83 (.012)	.37	0.94 (.125)	-1.02 (.121)	1.04 (.125)	-1.93 (.121)
G4i9e_G5i8e		.34 (.016)	.63	2.22 (.207)	1.58 (.142)	2.47 (.207)	1.18 (.142)
G3G4i10_G5i9		.58 (.016)	.53	1.19 (.115)	0.66 (.091)	1.32 (.115)	-0.46 (.091)
G4i11_G5i10		.56 (.017)	.56	1.31 (.122)	0.78 (.094)	1.46 (.122)	-0.35 (.094)
G5i11		.31 (.015)	.63	2.44 (.237)	1.70 (.168)	2.70 (.237)	1.60 (.168)
G3i5_G4G5i12		.40 (.016)	.38	0.75 (.090)	1.70 (.079)	0.83 (.090)	0.48 (.079)
G4G5i13a		.36 (.016)	.59	1.81 (.164)	1.58 (.120)	2.01 (.164)	0.95 (.120)
G4G5i13b		.75 (.014)	.46	1.09 (.120)	-0.22 (.107)	1.21 (.120)	-1.39 (.107)
G4G5i13c		.22 (.014)	.57	2.34 (.246)	2.09 (.210)	2.59 (.246)	2.42 (.210)
G3i8_G4G5i14		.76 (.014)	.54	1.53 (.154)	-0.08 (.135)	1.70 (.154)	-1.73 (.135)
G4G5i15		.35 (.016)	.40	0.81 (.094)	1.93 (.083)	0.90 (.094)	0.72 (.083)
G4G5i16		.48 (.017)	.57	1.38 (.126)	1.13 (.095)	1.53 (.126)	0.11 (.095)
G3i12_G4G5i17		.62 (.016)	.48	1.01 (.106)	0.43 (.088)	1.12 (.106)	-0.63 (.088)
G3i13a_G4G5i18a		.82 (.013)	.37	0.90 (.120)	-0.95 (.115)	1.00 (.120)	-1.80 (.115)
G3i13b_G4G5i18b		.74 (.015)	.62	2.12 (.168)	0.16 (.209)	2.35 (.209)	-1.89 (.168)
G3i13c_G4G5i18c		.63 (.016)	.49	1.07 (.091)	0.38 (.110)	1.19 (.110)	-0.72 (.091)
G3i13d_G4G5i18d		.73 (.015)	.55	1.48 (.123)	0.08 (.146)	1.64 (.146)	-1.43 (.123)

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

H.2. Spaghetti Plots

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

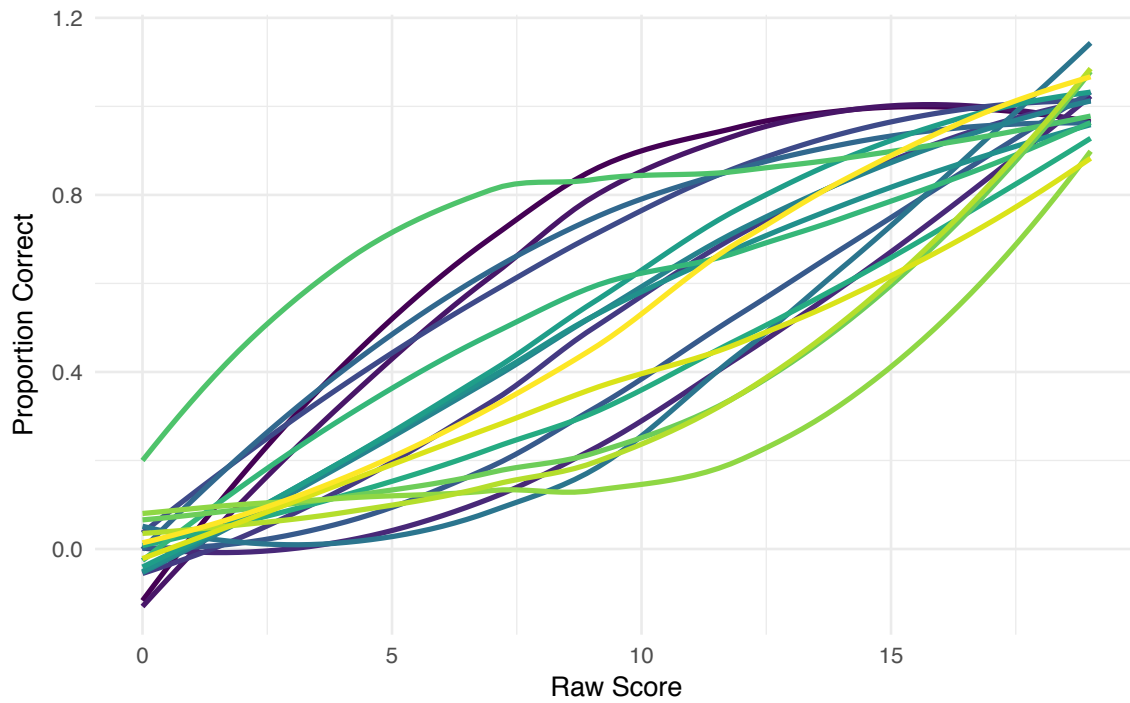


Figure H.1. Grade 3 spaghetti plot.

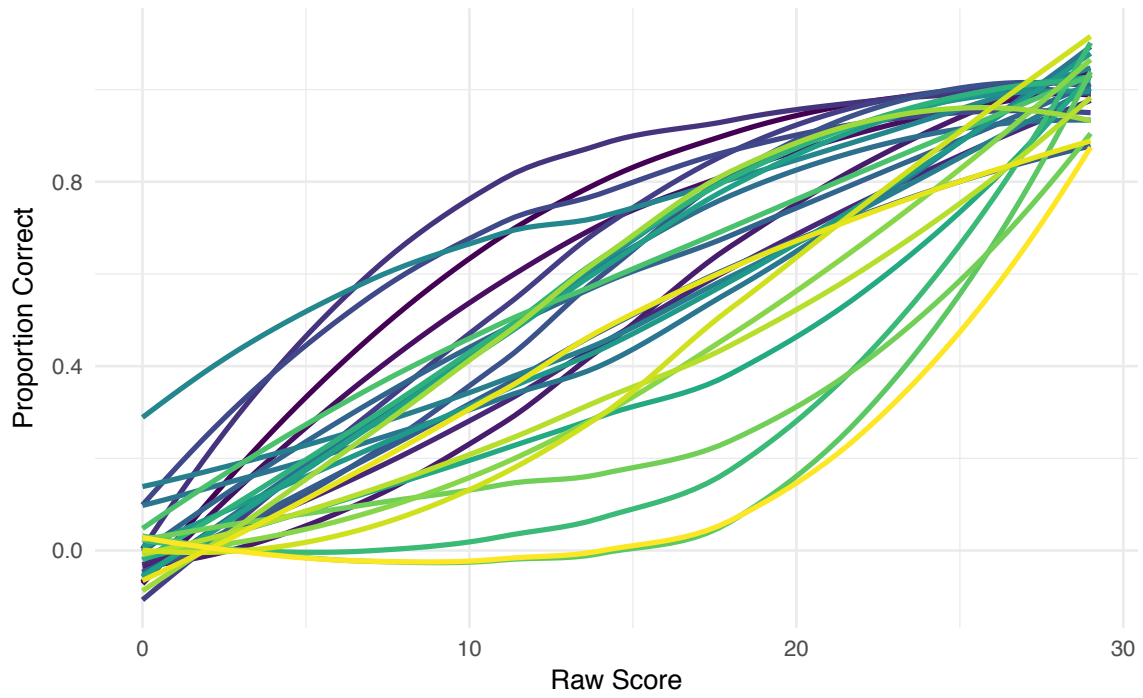


Figure H.2. Grade 4 spaghetti plot.

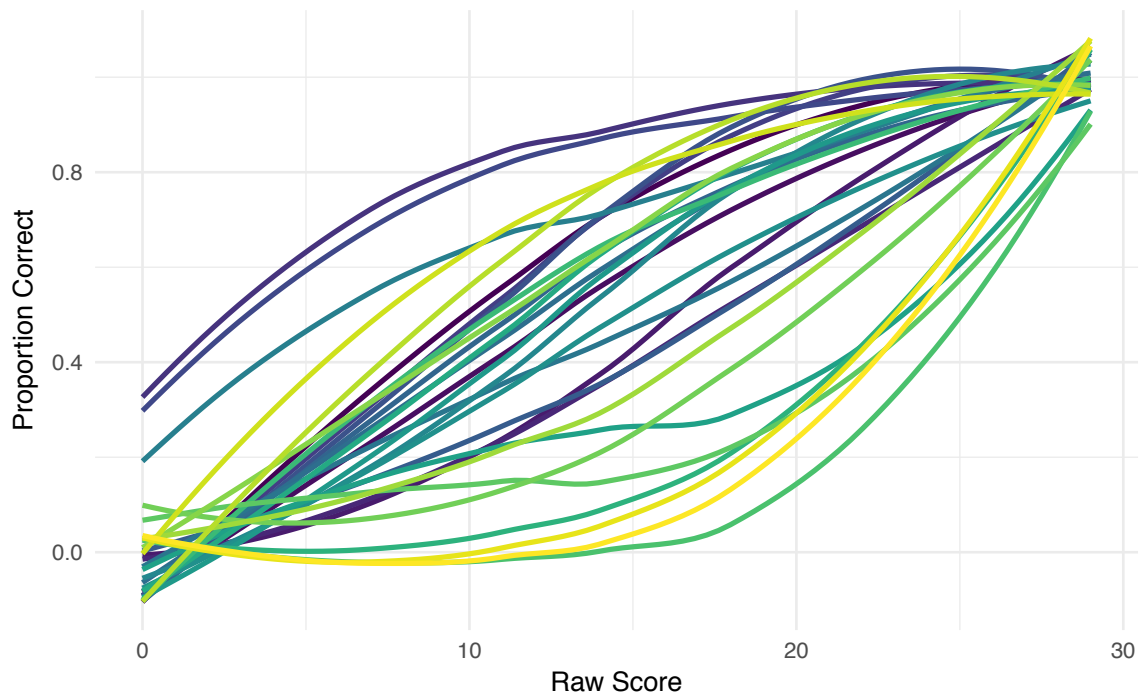


Figure H.3. Grade 5 spaghetti plot.

Appendix I. Most Common Incorrect Responses for Each Item

Table I.1. Proportion of Grade 3 Student Responses by Item

Item	Item description	Response format	Correct response	Most frequent incorrect responses			
			Response (%)	Response (%)	Response (%)	Response (%)	Response (%)
G3G4G5i1a		CR	3 (.53)	4 (.28)	2 (.08)	DNS (.03)	1 (.03)
G3G4G5i1b		CR	3 (.44)*	-	-	-	-
G3G4G5i2a		CR	$\frac{3}{4}$ (.11)	1 (.25)	3 (.10)	DNS (.08)	4 (.07)
G3G4G5i2b		CR	$\frac{3}{4}$ (.24)*	-	-	-	-
G3G4G5i3		CR	$\frac{1}{2}$ (.48)*	-	-	-	-
G3G4G5i4		CR	$\frac{1}{3}$ (.15)	1 (.28)	$\frac{1}{2}$ (.11)	2 (.09)	3 (.07)
G3i5_G4G5i12		CR	$\frac{3}{4}$ (.22)	$\frac{7}{8}$ (.25)	UI (.17)	1 (.10)	DNS (.08)
G3G4G5i6		SR	1 (.51)	$\frac{5}{6}$ (.48)	DNS (.01)	-	-
G3G4G5i7		SR	$\frac{1}{2}$ (.12)	$\frac{1}{10}$ (.79)	$\frac{1}{5}$ (.03)	$\frac{1}{9}$ (.03)	$\frac{1}{4}$ (.02)
G3i8_G4G5i14		SR	$\frac{1}{4}$ (.28)	$\frac{1}{2}$ (.31)	1 (.16)	$\frac{3}{4}$ (.11)	$\frac{3}{8}$ (.11)
G3G4i9a_G5i8a		CR	46 (.30)	56 (.06)	50 (.06)	DNS (.05)	194 (.04)
G3G4i9b_G5i8b		CR	6 (.31)	16 (.07)	DNS (.07)	5 (.04)	194 (.03)
G3G4i10_G5i9		CR	4 (.30)	19 (.14)	11 (.12)	3 (.06)	DNS (.05)
G3i11_G4i8		SR	$2\frac{1}{4}$ (.22)	$\frac{4}{9}$ (.27)	$1\frac{1}{2}$ (.18)	$\frac{1}{2}$ (.16)	3 (.15)
G3i12_G4G5i17		SR	12 (.38)	4 (.32)	8 (.13)	3 (.09)	1 (.04)
G3i13a_G4G5i18a		SR	$\frac{4}{5}$ (.74)	$\frac{3}{5}$ (.14)	Both (.06)	DNS (.05)	UI (.01)
G3i13b_G4G5i18b		SR	$\frac{3}{5}$ (.19)	$\frac{3}{7}$ (.70)	DNS (.05)	Both (.05)	UI (.01)
G3i13c_G4G5i18c		SR	$\frac{5}{4}$ (.17)	$\frac{6}{7}$ (.74)	DNS (.06)	Both (.03)	UI (<.01)
G3i13d_G4G5i18d		SR	$\frac{2}{3}$ (.17)	$\frac{5}{12}$ (.76)	DNS (.06)	Both (.01)	-

Note. $n = 1,045$ valid grade 3 tests conducted. Boldface 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." Items with asterisks were entered as correct or incorrect.

Table 1.2. Proportion of Grade 4 Student Responses by Item

Item	Item description	Response format	Correct response	Most frequent incorrect responses			
			Response (%)	Response (%)	Response (%)	Response (%)	Response (%)
G3G4G5i1a		CR	3 (.67)	4 (.20)	2 (.06)	DNS (.02)	5 (.01)
G3G4G5i1b		CR	3 (.59)*	-	-	-	-
G3G4G5i2a		CR	$\frac{3}{4}$ (.37)	1 (.19)	$\frac{1}{6}$ (.06)	DNS (.05)	3 (.04)
G3G4G5i2b		CR	$\frac{3}{4}$ (.39)*	-	-	-	-
G3G4G5i3		CR	$\frac{1}{2}$ (.77)*	-	-	-	-
G3G4G5i4		CR	$\frac{1}{3}$ (.55)	1 (.12)	$\frac{1}{2}$ (.08)	$\frac{1}{4}$ (.04)	$\frac{1}{1}$ (.03)
G4G5i5		CR	$\frac{3}{7}$ (.51)	3 (.11)	$\frac{3}{4}$ (.05)	$\frac{1}{3}$ (.05)	4 (.04)
G3G4G5i6		SR	1 (.71)	$\frac{5}{6}$ (.29)	DNS (<.01)	-	-
G3G4G5i7		SR	$\frac{1}{2}$ (.48)	$\frac{1}{10}$ (.49)	$\frac{1}{4}$ (.01)	$\frac{1}{9}$ (.01)	$\frac{1}{5}$ (.01)
G3i11_G4i8		SR	$2\frac{1}{4}$ (.41)	$\frac{4}{9}$ (.24)	$1\frac{1}{2}$ (.18)	3 (.08)	$\frac{1}{2}$ (.08)
G3G4i9a_G5i8a		CR	46 (.50)	56 (.06)	36 (.05)	DNS (.05)	50 (.04)
G3G4i9b_G5i8b		CR	6 (.51)	5 (.07)	16 (.07)	DNS (.06)	14 (.03)
G3G4i9c_G5i8c		CR	$\frac{3}{7}$ (.29)	$\frac{3}{21}$ (.21)	24 (.09)	DNS (.08)	$\frac{1}{21}$ (.03)
G3G4i9d_G5i8d		CR	$\frac{2}{4}$ (.39)	$\frac{2}{0}$ (.12)	2 (.08)	DNS (.08)	$\frac{4}{7}$ (.05)
G3G4i9e_G5i8e		CR	$\frac{3}{10}$ (.05)	$\frac{2}{12}$ (.37)	14 (.08)	DNS (.08)	$\frac{1}{12}$ (.06)
G3G4i10_G5i9		CR	4 (.41)	19 (.09)	11 (.08)	3 (.04)	60 (.04)
G4i11_G5i10		CR	1 (.29)	0 (.09)	DNS (.08)	$\frac{1}{3}$ (.07)	$\frac{3}{6}$ (.06)
G3i5_G4G5i12		CR	$\frac{3}{4}$ (.28)	$\frac{7}{8}$ (.23)	UI (.15)	$\frac{3}{8}$ (.14)	DNS (.08)
G4G5i13a		CR	$\frac{9}{6}$ (.12)	$\frac{1}{2}$ (.18)	$\frac{1}{3}$ (.15)	DNS (.09)	$\frac{1}{4}$ (.06)
G4G5i13b		CR	2 (.52)	DNS (.09)	$\frac{1}{3}$ (.03)	$\frac{1}{2}$ (.03)	3 (.03)
G4G5i13c		CR	$\frac{17}{6}$ (.06)	DNS (.10)	$\frac{2}{5}$ (.08)	3 (.05)	$\frac{5}{6}$ (.05)
G3i8_G4G5i14		SR	$\frac{1}{4}$ (.53)	$\frac{1}{2}$ (.20)	$\frac{3}{4}$ (.10)	1 (.08)	$\frac{3}{8}$ (.05)
G4G5i15		SR	D (.20)	B (.32)	C (.31)	A (.07)	E (.06)
G4G5i16		SR	B (.28)	D (.36)	E (.14)	C (.10)	A (.08)
G3i12_G4G5i17		SR	12 (.44)	4 (.24)	3 (.12)	8 (.11)	DNS (.05)
G3i13a_G4G5i18a		SR	$\frac{4}{5}$ (.71)	$\frac{3}{5}$ (.19)	DNS (.07)	Both (.02)	UI (.01)
G3i13b_G4G5i18b		SR	$\frac{3}{5}$ (.52)	$\frac{3}{7}$ (.38)	DNS (.07)	Both (.02)	UI (.01)
G3i13c_G4G5i18c		SR	$\frac{5}{4}$ (.41)	$\frac{6}{7}$ (.44)	DNS (.07)	Both (.06)	UI (.01)
G3i13d_G4G5i18d		SR	$\frac{2}{3}$ (.51)	$\frac{5}{12}$ (.39)	DNS (.08)	Both (.01)	UI (<.01)

Note. $n = 663$ valid grade 4 tests conducted. Items that were not answered were recorded as “DNS” Item responses that were unclear were recorded as “UI.” Items with asterisks were entered as correct or incorrect.

Table I.3. Proportion of Grade 5 Student Responses by Item

Item	Item description	Response format	Correct response	Most frequent incorrect responses				
			Response (%)	Response (%)	Response (%)	Response (%)	Response (%)	
G3G4G5i1a		CR	3 (.80)	4 (.10)	2 (.05)	12 (.01)	DNS (.01)	
G3G4G5i1b		CR	3 (.71)*	-	-	-	-	
G3G4G5i2a		CR	$\frac{3}{4}$ (.63)	1 (.12)	DNS (.04)	$\frac{1}{6}$ (.03)	$\frac{2}{4}$ (.03)	
G3G4G5i2b		CR	$\frac{3}{4}$ (.56)*	-	-	-	-	
G3G4G5i3		CR	$\frac{1}{2}$ (.92)*	-	-	-	-	
G3G4G5i4		CR	$\frac{1}{3}$ (.82)	$\frac{1}{2}$ (.05)	$\frac{1}{4}$ (.03)	1 (.03)	3 (.01)	
G4G5i5		CR	$\frac{3}{7}$ (.78)	3 (.03)	$\frac{3}{4}$ (.03)	$\frac{1}{3}$ (.02)	$\frac{4}{7}$ (.02)	
G3G4G5i6		SR	1 (.90)	$\frac{5}{6}$ (.10)	DNS (<.01)	-	-	
G3G4G5i7		SR	$\frac{1}{2}$ (.82)	$\frac{1}{10}$ (.16)	$\frac{1}{4}$ (<.01)	$\frac{1}{9}$ (<.01)	$\frac{1}{5}$ (<.01)	
G3G4i9a_G5i8a		CR	46 (.74)	56 (.04)	36 (.03)	DNS (.02)	50 (.02)	
G3G4i9b_G5i8b		CR	6 (.74)	16 (.06)	5 (.04)	DNS (.03)	15 (.01)	
G3G4i9c_G5i8c		CR	$\frac{3}{7}$ (.85)	$\frac{3}{21}$ (.06)	$\frac{1}{21}$ (.01)	DNS (.01)	24 (.01)	
G3G4i9d_G5i8d		CR	$\frac{2}{4}$ (.83)	$\frac{2}{0}$ (.03)	2 (.02)	DNS (.02)	$\frac{4}{4}$ (.01)	
G3G4i9e_G5i8e		CR	$\frac{6}{10}$ (.34)	$\frac{2}{12}$ (.24)	$\frac{2}{10}$ (.15)	$\frac{1}{12}$ (.04)	DNS (.03)	
G3G4i10_G5i9		CR	4 (.58)	60 (.09)	11 (.06)	3 (.05)	19 (.03)	
G4i11_G5i10		CR	1 (.56)	$\frac{1}{3}$ (.18)	0 (.06)	$\frac{3}{3}$ (.05)	$\frac{1}{2}$ (.02)	
G5i11		CR	$3\frac{1}{4}$ (.31)	$2\frac{4}{6}$ (.12)	3 (.09)	$2\frac{4}{4}$ (.03)	DNS (.03)	
G3i5_G4G5i12		CR	$\frac{3}{4}$ (.40)	$\frac{7}{8}$ (.23)	$\frac{3}{8}$ (.16)	$\frac{5}{8}$ (.10)	UI (.04)	
G4G5i13a		CR	$\frac{9}{6}$ (.36)	$\frac{1}{2}$ (.15)	$\frac{1}{3}$ (.07)	DNS (.04)	$\frac{1}{4}$ (.03)	
G4G5i13b		CR	2 (.75)	DNS (.04)	$\frac{1}{2}$ (.02)	$1\frac{1}{2}$ (.01)	$\frac{1}{6}$ (.01)	
G4G5i13c		CR	$\frac{17}{6}$ (.22)	$\frac{2}{5}$ (.06)	DNS (.05)	$\frac{5}{6}$ (.04)	$\frac{2}{3}$ (.04)	
G3i8_G4G5i14		SR	$\frac{1}{4}$ (.76)	$\frac{3}{4}$ (.08)	$\frac{1}{2}$ (.07)	1 (.04)	$\frac{3}{8}$ (.03)	
G4G5i15		SR	D (.35)	B (.29)	C (.22)	E (.06)	A (.06)	
G4G5i16		SR	B (.48)	D (.21)	E (.12)	C (.09)	A (.08)	
G3i12_G4G5i17		SR	12 (.62)	4 (.15)	3 (.11)	8 (.08)	1 (.03)	
G3i13a_G4G5i18a		SR	$\frac{4}{5}$ (.82)	$\frac{3}{5}$ (.13)	Both (.03)	DNS (.02)	UI (<.01)	
G3i13b_G4G5i18b		SR	$\frac{3}{5}$ (.74)	$\frac{3}{7}$ (.20)	Both (.03)	DNS (.03)	UI (<.01)	
G3i13c_G4G5i18c		SR	$\frac{5}{4}$ (.63)	$\frac{6}{7}$ (.27)	Both (.05)	DNS (.03)	UI (.02)	
G3i13d_G4G5i18d		SR	$\frac{2}{3}$ (.73)	$\frac{5}{12}$ (.20)	Both (.05)	DNS (.02)	UI (.01)	

Note. $n = 906$ valid grade 5 tests conducted. Items that were not answered were recorded as "DNS" Item responses that were unclear were recorded as "UI." Items with asterisks were entered as correct or incorrect.