

Florida State University Libraries

2017

Psychometric Report for the Early Fractions Test Administered with Third- and Fourth-grade Students in Fall 2016

Robert C Schoen, Sicong Liu, Xiaotong Yang and Insu Paek



Psychometric Report for the Early Fractions Test Administered with Third- and Fourth-grade Students in Fall 2016

Robert C. Schoen
Sicong Liu
Xiaotong Yang
Insu Paek

AUGUST 2017

Research Report No. 2017-10

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A150043 to Mills College. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Suggested citation: 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.

Copyright 2017, Florida State University. All rights reserved. Requests for permission to use these materials should be directed to Robert Schoen, rschoen@lsi.fsu.edu, FSU Learning Systems Institute, 4600 University Center C, Tallahassee, FL, 32306.

Psychometric Report for the Early Fractions Test Administered with Third- and Fourth-grade Students in Fall 2016

Research Report No. 2017-10

Robert C. Schoen

Sicong Liu

Xiaotong Yang

Insu Paek

August 2017

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

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

Robert Schoen directed the data collection and report-writing processes and assisted in guiding and interpreting the analytic methods and results. Sicong Liu and Xiaotong Yang contributed equally to the data analysis and IRT model calibration as well as the writing of data analysis and results sections of the report. Insu Paek provided overall guidance for the data modeling and scoring and provided guidance and feedback on the various drafts of the report.

Catherine Lewis, Rebecca Perry, and Kevin Lai developed the test, primarily through selection or adaptation of items drawn from other published sources. Robert Schoen, Claire Riddell, and several members of the advisory board for the larger project, including Akihiko Takahashi, Tad Watanabe, Phil Daro, and Geoffrey Saxe reviewed the test items and provided feedback. Claire Riddell managed the distribution and collection of tests and consent forms for students. Kristy Farina managed the data entry and verification process. Along with Claire and Kristy, Shelby McCrackin and Alex Utecht assisted with data entry, verification of accuracy, and adjudication.

Catherine Lewis, Kevin Lai, Amanda Tazaz, and Charity Bauduin reviewed the final draft and provided useful feedback to improve the report. Any remaining errors or shortcomings are the responsibility of the authors.

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

Table of Contents

Acknowledgements.....	iv
Executive Summary.....	1
Purpose Statement	1
Description of the Test.....	1
Sample and Setting	1
Results.....	1
Item Diagnostics and Scoring.....	1
Dimensionality	2
IRT Data Modeling.....	2
Reliability.....	2
Distribution of Student Ability Scores.....	2
Evidence of External Validity.....	2
Discussion and Conclusions	3
1. Introduction	4
2. Initial Item Review	6
3. Data and Scoring.....	8
3.1. Sample.....	8
3.2. Data Entry and Verification Procedures.....	9
3.3. Item Scoring	9
4. Dimensionality Analysis	12
4.1. Exploratory Factor Analysis.....	12
4.2. Parallel Analysis.....	13
5. Classical Testing Theory (CTT) Analyses.....	14
5.1. Distribution of the Observed Test Score.....	14
5.2. Item Difficulty & Discrimination.....	14
5.3. Reliability & Standard Error of Measurement	16
6. Item Response Theory (IRT) Analyses.....	17
6.1. Model Description.....	17
6.2. Item Difficulty and Discrimination	18

6.3. Test Information and Estimated Person Ability 21

7. Additional Analyses..... 24

 7.1. Intraclass Correlation Coefficient..... 24

 7.2. Predictive Validity 24

8. Discussion and Conclusions 25

References 27

Appendix A. The Early Fractions Test Form 29

Appendix B. Administration Instructions 39

Appendix C. Scoring Key..... 41

List of Tables

Table 1.1. Test Blueprint for the Original Test Form and the Final Scale	4
Table 2.1. Detailed Test Blueprint for the Fall 2016 Early Fractions Test, Split by Phase in Data Analysis ..	7
Table 3.1. Demographic Characteristics of the Students (n = 1,400) in the Fall 2016 Field-test of the Early Fractions Test.....	8
Table 3.2. Item Indexing and Scoring for both Test-Form and Final-Scale Format	11
Table 4.1. Eigenvalues Estimated from Mplus and Their Corresponding Percentages of Explained Variation.....	12
Table 5.1. Item Difficulty and Discrimination from CTT Analyses.....	15
Table 6.1. Descriptive Statistics of Discrimination Index and Difficulty Index of all the 18 Items.....	18
Table 6.2. Parameter Estimates and Standard Errors for Final-Scale Items Modeled Using 2PL.....	19
Table 6.3. Parameter Estimates and Standard Errors for Final-Scale Items Modeled using 3PL	19
Table 6.4. Parameter Estimates and Standard Errors for Final-Scale Items Modeled Using GPCM.....	19

List of Figures

Figure 4.1. Scree plot of eigenvalues estimated from Mplus.	13
Figure 5.1. Bar graph depicting the distribution of the observed test score in the final-scale format.	14
Figure 6.1. Item discrimination estimate (a) of each final-scale item.	20
Figure 6.2. Item difficulty estimate (b) of each final-scale item.	20
Figure 6.3. Test information curve and conditional standard errors of measurement (CSEM) for the final-scale format.	21
Figure 6.4. Person abilities (i.e., θ) estimated by maximum likelihood estimation (MLE).	23
Figure 6.5. Person abilities (i.e., θ) estimated by expected a posteriori (EAP).....	23

List of Equations

Equation 1. Item Difficulty Index from CTT Analyses.....	15
Equation 2. Standard Error of Measurement (SEM) from CTT Analyses	16
Equation 3. Two-Parameter (2PL) Model	17
Equation 4. Three-Parameter (3PL) Model.....	17
Equation 5. Generalized Partial Credit Model (GPCM).....	18
Equation 6. Conditional Standard Error of Measurement (CSEM) Given Person Ability.....	21

Executive Summary

The Early Fractions Test is a paper-pencil test designed to measure mathematics achievement of third- and fourth-grade students in the domain of fractions. The test was administered to a sample of 1,400 third- and fourth-grade students as part of a larger study involving a multisite cluster randomized trial evaluation design to investigate the effects of lesson study and a fractions resource toolkit on classroom instruction and student achievement in fractions.

Purpose Statement

The purpose, or intended use, of the Early Fractions Test is to serve as a student pretest covariate and a test of baseline equivalence in the larger study. In this report, we discuss our exploration of options for scoring and data modeling and make recommendations for optimal scoring and data modeling procedures. We also report on the results of data modeling, including analyses of dimensionality, scale reliability estimates, the intraclass correlation coefficient for the 66 schools represented in the sample, and the percentage of the variance in student achievement as measured by the end-of-year mathematics test that is explained by their scores on this beginning-of-year test.

Description of the Test

The Early Fractions Test is designed to measure the competence of third- and fourth-grade students in early fractions. The content is designed to align with the Common Core State Standards for Mathematics and a related intervention involving lesson study with a fractions resource toolkit (Lewis & Perry, 2017). The test form contains 20 numbered items prompting up to 34 responses from the test taker, with nine of them using a selected-response format and 25 of them using a constructed-response format. Each of the 34 responses was scored dichotomously (i.e., correct, incorrect) in accordance with a scoring key provided by the test developers.

Sample and Setting

The Early Fractions Test was administered with a sample of 1,400 third- and fourth-grade students in six U.S. states in fall 2016. A single test form was used with all the students in the sample. The teachers of the students in the sample were participating in a large-scale randomized controlled trial of lesson study with a fractions resource toolkit.

Results

Item Diagnostics and Scoring

Item diagnostics and calibration accounting resulted in the collapsing of the 34 individual responses (or non-responses) to a total of 18 independent items. All the 34 responses contribute to the final 18-item scale. Initial screening of the items used an approach based on classical test theory. Item difficulty indices for the 18 items in the final scale ranged from .12 to .81. The lowest item-rest correlation coefficient was .29. All the other items had item-rest correlation coefficients greater than .38, suggesting that the items generally had good discriminative power.

Dimensionality

To investigate the dimensionality of the test data, we performed Exploratory Factor Analysis and Parallel Analysis. The results of these analyses both suggested a single dominant factor and supported an assumption of unidimensionality in the data.

IRT Data Modeling

Because the test form contained a mix of selected-response and constructed-response items resulting in dichotomous and polytomous variables, the data were modeled with a combination of a 2-parameter logistic model, a 3-parameter logistic model (to adjust for student guessing), and a Generalized Partial Credit Model. The models are based on item-response theory (IRT). They were run using flexMIRT (version 3.5) software (Cai, 2017). Maximum likelihood estimator and *expected a posteriori* estimator were used in calculating the person ability estimates. A maximum likelihood estimator is generally supported for estimating person ability in educational testing. However, due to computational reasons, it cannot provide person ability estimates for students who have perfect or zero test scores (de Ayala, 2009). To help estimate these extreme cases, we also used *expected a posteriori* estimator.

Reliability

Using a classical test theory approach, Coefficient α and standard error of measurement were calculated to be .85 and 2.74, respectively. Additional information of test information and conditional standard error of measurement was generated through the IRT approach. Test information and the conditional standard error of measurement (CSEM) are inversely related (see Formula 6). Figure 6 displays the test information curve and CSEM, suggesting that the highest test information and the lowest CSEM occurred when the person ability (i.e., θ) was approximately 0.80. Also, the person ability estimate was most reliable (i.e., lowest CSEM) for the person ability between -0.80 and 1.60 on the θ scale, but was least reliable (i.e., highest CSEM) for the person ability less than -2.00 on the θ scale.

Distribution of Student Ability Scores

Using an *expected a posteriori* (EAP) technique, we found that the distribution of student ability (θ) scores for the third- and fourth-grade students in the present sample does not appear to be much different from a normal distribution. The sample distribution of θ scores resulting from the EAP for the 1,400 third- and fourth-grade students ranged from -1.95 to 2.67 with a mean of 0.00 and standard deviation of 0.94 .

Based on the sample data from 1,400 grade 3 or 4 students representing 66 schools, we calculated a school-level intraclass correlation coefficient (ICC) of .45 using the person ability (theta) estimates generated by the EAP estimator.

Evidence of External Validity

One of the two primary intended uses of the Early Fractions Test scores is to serve as a pretest student achievement covariate in models examining the contrast between school mathematics achievement in schools in the treatment and comparison groups of a randomized controlled trial.

Lewis and Perry (2017) used an almost identical test that was scored using a classical test theory approach. Results reported by Lewis and Perry provide evidence that the test may be sufficiently sensitive to detect a treatment effect from the same intervention being implemented in the larger study

from which the present data were used. Similar analyses and results are not available for the larger study at the time of publication of the present report.

To examine the potential strength of the covariate based on the person ability (theta) estimates generated by the Early Fractions Test, we calculated the correlation between the ability estimates generated by the pre- and posttest student scores. Based on a sample of 1,134 students who completed both the pretest and the posttest, and using SPSS version 22, we found a Pearson correlation coefficient of .69 between the ability (theta) estimates at pretest using the EAP estimator and posttest ability estimates. Using the EAP estimator, with no adjustment for other factors such as clustering in schools, the student ability estimates from the Early Fractions Test used at the beginning of the school year explains 47% of the variance in student scores (i.e., $R^2 = .47$) as measured at the end of the school year.¹

Discussion and Conclusions

Several of the responses in the Early Fractions Test involved item sets, which present a potential threat to the validity of an assumption of local-independence. We found evidence of collinearity of items in item sets when the items were modeled as separate items scored dichotomously. We also found that collapsing those responses into polytomous variables preserved the assumption of the local-independence assumption and resulted in items with stable parameter estimates. The Early Fractions Test appears to be measuring a single, dominant trait, supporting an assumption of unidimensionality in the data. Evaluation of the structural validity of the resulting 18-item scale supports the assertion that the Early Fractions Test meets or exceeds standards for educational and psychological measurement.

¹ We note that the pretest and posttest scores were not equated.

1. Introduction

The Early Fractions Test is designed to measure student understanding of early fractions concepts, including awareness of the referent unit (or whole), partitioning the referent unit into unit fractions and iterating unit fractions to compose non-unit fractions, fractions as corresponding to points (or numbers) on a number line, magnitude of fractions, and operations on fractions. Many items require students to be familiar with conventional terminology and notation for common fractions (e.g., one-sixth, $3\frac{1}{2}$). Items do not involve decimal numbers (e.g., 0.17, 3.50). These topics are aligned with the content of the third- and fourth-grade standards in the Common Core State Standards for Mathematics (CCSS-M; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010).

Table 1.1. Test Blueprint for the Original Test Form and the Final Scale

Category	Number of items	
	Test form	Final scale
Fractions as Number on a Number Line	4	4
Magnitude Comparison	2	2
Partitioning and Iterating	8	8
Operations on Fractions	3	1
Referent Unit	3	3
<i>Total</i>	<i>20</i>	<i>18</i>

Table 1.1 shows the test blueprint as seen on the test form by the students as well as a test blueprint corresponding to the final scale after items in testlets were transformed into polytomous items as described in subsequent sections of this report. The blueprint shows how the test questions relate to the learning standards in accordance with the following categories within the domain of fractions knowledge: *referent unit, partitioning and iterating, fractions as number on a number line, magnitude comparison, and operations on fractions.*

The purpose, or intended use, of the Early Fractions Test is to serve as a student pretest covariate and a test of baseline equivalence in a larger study focused on evaluating the impact of an educational intervention on student learning. The larger study involves a two-by-two factorial design in a multisite cluster randomized trial designed to investigate the effects of lesson study and a fractions resource toolkit on classroom instruction and student learning in fractions.

The present report focuses on scoring and data modeling based on the data generated by the Early Fractions Test. The Early Fractions Test was administered with a sample of 1,400 third- and fourth-grade students in fall 2016. Grades 3 and 4 students completed the same test form. The purpose of the present report is to describe the item- and scale-scoring procedures using data from the administration of the fall 2016 student pretest for the randomized controlled trial and provide an evaluation of available evidence supporting the substantive, structural, and external validity (Flake, Pek, & Hehman, 2017) of the test scores.

Lewis and Perry (2017) used a previous version of the Early Fractions Test in their evaluation of lesson study with a fractions resource toolkit. The previous version and this version both drew from released items from U.S. state and national assessments, published curricula, and research articles (Beckmann,

2005; California Department of Education, n.d.; Hackenberg, Norton, Wilkins, & Steffe, 2009; Hironaka & Sugiyama, 2006; National Assessment of Educational Progress, 1992; Van de Walle, 2007).

The current version of the Early Fractions Test was modified by the senior personnel on a research team conducting a subsequent randomized controlled trial evaluating the impact of lesson study and fractions resource kits. Several items were modified to clarify the instructions to the respondent, and several other items involving symbolic computation and understanding of equipartitioning were drawn from a researcher-created test designed to measure student understanding of early fractions knowledge aligned with the CCSS-M (Schoen, Anderson, Riddell, & Bauduin, 2017).

2. Initial Item Review

The Early Fractions Test requires students to make 34 fraction-related responses that are indexed into 20 items on the form the students saw when they took the test. The test form contains 20 numbered items prompting up to 34 responses from the test taker. Nine of these responses involve a selected-response format, and 25 of them involve a constructed-response format. The discrepancy between 34 and 20 exists, because several items (i.e., items 1, 2, 10, 11, 12, 16 as they are enumerated on the form) require more than one response from students. During data entry, the 34 responses are initially coded into 34 dichotomous variables. Dichotomous variables indexed under the same test item are then added together to form a polytomous variable to represent the item. In the end, the 34 response variables are recoded into 20 item variables.

Because we use item response theory models in scoring students' latent ability, the recoding was performed in an effort to address concerns about local dependence of items. Subsequently, based on statistical reasons explained in section 3.3 of this report, the 20 item variables were again recoded into 18 item variables after combining items 7, 8, and 9 into a single item. In attempt to clarify our references to items in this report, we label the 34, 20, and 18 coding format using the description of data-entry, test-form, and final-scale, respectively. We also differentiated test-form items and final-scale items by assigning an "*" after each final-scale item number. For example, item 1 represents the first item of the test-form format, and item 1* stands for the first item of the final-scale format.

Although the description of data analysis is presented in a linear-sequential fashion in this report, the analyses were completed through an interactive, overlapping, and iterative process. For instance, the decision to recode the 20 test-form variables into the 18 final-scale variables was informed by the polychoric correlations between the items, and the item discrimination index provided from the Item-Response Theory analysis. Table 2.1 provides a detailed blueprint for the test and includes a map of the correspondence among the data-entry, test-form, and final-scale formats.

Table 2.1. Detailed Test Blueprint for the Fall 2016 Early Fractions Test, Split by Phase in Data Analysis

Question description	Test form #	Data entry #	Final scale #
Fractions as Number on a Number Line			
Rabbit Problem Part 1	1	1a	1*
Rabbit Problem Part 2	1	1b	1*
Polar Bear Problem Part 1	2	2a	2*
Polar Bear Problem Part 2	2	2b	2*
Mark $\frac{3}{4}$ on a NL	15	15	13*
Determine $\frac{9}{8}$ on NL	16	16a	14*
Determine 2 on NL	16	16b	14*
Determine $\frac{22}{8}$ on NL	16	16c	14*
Magnitude Comparison			
1 gallon vs. $\frac{5}{6}$ gallon (Pretest Open Response/Posttest Circle)	5	5	5*
Determining the greatest fraction (Pretest MC)	6	6	6*
Partitioning and Iterating			
Part of a Referent Unit ($\frac{2}{3}$)	3	3	3*
Partitioned $\frac{1}{6}$	4	4	4*
Equal Partitioning (fourths) Shape 1- irregular	10	10a	8*
Equal Partitioning (fourths) Shape 2- horizontal rectangle fourths	10	10b	8*
Equal Partitioning (fourths) Shape 3- vertical rectangle fourths	10	10c	8*
Equal Partitioning (fourths) Shape 4- circle	10	10d	8*
Equal Partitioning (fourths) Shape 5- vertical rectangle thirds	10	10e	8*
$\frac{1}{3}$ of the shaded ribbon	11	11a	9*
Shade $\frac{1}{2}$	11	11b	9*
Shade $\frac{3}{4}$	11	11c	9*
Shade $\frac{5}{6}$	11	11d	9*
Iterating unit fraction box (3 pieces of $\frac{1}{4}$ is box/box)	12	12a	10*
Iterating unit fraction box (3 pieces of $\frac{1}{5}$ is $\frac{3}{5}$)	12	12b	10*
Iterating unit fraction box (box pieces of $\frac{1}{10}$ is $\frac{7}{10}$)	12	12c	10*
Iterating unit fraction box (box/8 = 1)	12	12d	10*
Fourths in a whole	13	13	11*
Fourths in 3	14	14	12*
Joe's walk	20	20	18*
Operations on Fractions			
	9	9	7*
	7	7	7*
	8	8	7*
Referent Unit			
Jose and Ella's pizzas	17	17	15*
Determining Referent Unit from $\frac{3}{5}$	18	18	16*
Draw $\frac{4}{3}$	19	19	17*
Total Number of Items	20	34	18

Note. Question Description = description of the fraction questions; Test Form # = the index numbers of all the items in the original fraction test (see Appendix A); Data Entry # = the index numbers of data entry (dichotomous) variables that correspond to all the 34 responses tapped by the test; Final Scale # = the adjusted index numbers (with an * behind to help differentiate from test-form item numbers) of all the items in the statistical analyses.

3. Data and Scoring

3.1. Sample

The Early Fractions Test was administered with 1,400 third- and fourth-grade students in six U.S. states in fall 2016 in a paper-pencil format. The students were enrolled in schools where teachers had volunteered to participate in a randomized-controlled trial designed to investigate the effects of lesson study and fractions resource toolkits on student learning. The sample mean proportion of students eligible for free or reduced-price lunch was .62, and the standard deviation was .23. The proportion of students in each school who were eligible for free or reduced-price lunch ranged from a minimum of .15 to a maximum of .98. Table 3.1 provides descriptive statistics describing the available characteristics of the analytic sample.

Table 3.1. Demographic Characteristics of the Students (n = 1,400) in the Fall 2016 Field-test of the Early Fractions Test

Characteristic	Number (Proportion of sample)
Language	
Emergent bilingual	208 (.15)
Non-ELL	904 (.65)
Unknown	288 (.21)
Grade level	
Third	618 (.44)
Fourth	782 (.56)
Gender	
Male	573 (.41)
Female	591 (.42)
Unknown	236 (.17)
State	
FL	852 (.61)
CA	182 (.13)
IL	232 (.17)
NY	90 (.06)
CO	25 (.02)
IN	19 (.01)

Note. Gender and English-learner status were indicated by the students' classroom teachers. Other individual student demographic characteristics, such as ethnicity, exceptionality, or eligibility for free or reduced-price lunch, were not available at the time of writing the report. Some of the percentages do not sum to 1.00 due to rounding errors.

All students completed the same test form, which is provided in Appendix A. Test forms were mailed to participating schools by research project staff at Florida State University. The tests were administered by school employees, usually the students' classroom teachers. Administration instructions accompanied the tests and are provided in Appendix B. Administration of the tests occurred during a period spanning August 2016 through January 2017.

3.2. Data Entry and Verification Procedures

A team of four research assistants performed data entry in accordance with a detailed protocol. The data entry personnel were not informed of the assigned treatment condition of the participating schools. Test data were entered into a forms-based FileMaker database using item-specific data validation protocols. The students' responses were recorded as they were written for selected-response and fill-in-the-blank items. Other constructed-response items were scored during the data entry process according to the criteria set forth in the scoring rubric (provided in Appendix C), and only an indication of correct or incorrect was recorded for these items. Skipped items were scored as incorrect non-responses. As a result, there were no missing item-level data in the data set. Responses to fill-in-the-blank items were adjudicated by a committee that determined whether each response warranted a correct or incorrect score in accordance with the guidelines established by the scoring rubric.

To verify that data entry and scoring guidelines were being conducted consistently across data entry personnel, a random sample of seven schools (representing 11% of the total sample) was selected for double-entry. Data entry personnel were not informed when they were assigned a set of tests that were selected for double-entry. For this comparison, a second person entered the response data into the Filemaker system for the sampled students and entered them in a new data entry form. The two entries were scored separately as correct or incorrect as described in the preceding paragraph, and the scored data were compared for agreement between the two sets of data. The scored data agreed at a rate greater than 99% between the two records when compared on all items.

3.3. Item Scoring

The test developers provided an answer key and scoring rubric that was used to determine whether to judge responses as correct or incorrect. The scoring rubric is provided in Appendix C. Every attempt was made to score items drawn from external sources in a manner consistent with the item developers' intentions. For instance, one of the constructed-response items was drawn from a set of released items from the NAEP, and the NAEP rubric was used for scoring.

As explained previously in this report, 34 data-entry variables were recoded into 20 test-form variables (that correspond to the item indexing of the original test), because several test-form items (i.e. items 1, 2, 10, 11, 12, 16) require more than one response from students. For example, item 1 requires two responses, and item 11 requires four responses (see the items in the test form provided in Appendix A). To score these test-form items with more than one response, we formed polytomous variables consisting of the sum of the response scores. This resulted in a collapsing of the 34 individual responses to 20 items.

Although we scored each of the 20 test-form items in the last step, we further adjusted the item coding in two special cases based on statistical reasons. In the first case, items 7, 8 and 9 were three fill-in-the-blank items that were dichotomously scored. We combined these three items into one polytomous item (i.e., item 7*) based on the following analyses. First, the three items were placed together under the same instruction in the test (see Appendix A). This raises concerns about dependency in items. Second, the polychoric correlations between any two of these three items were very high (i.e., 0.99 for item 7 & 8, 1.00 for item 7 & 9, and 0.99 for item 8 & 9). This suggests that the three items are essentially providing evidence of a single ability dimension. Third, when the three items were assumed to be three independent items in models based on item response theory using flexMIRT 3.5 software (Cai, 2017),

the item discrimination estimates for these three items ranged between 6 and 20, which is unreasonably large. When we combined the three items into a single polytomous item (i.e., item 7*), the estimate of discrimination index became stable and produced a reasonable value (0.65). After combining the three items, the total number of independent items on the test scale was reduced from 20 in the test-form format to 18 in the final-scale format (see Table 1.1).

In the second case, item 10 (i.e., item 8*) requests up to five responses from the test taker. (See Appendix A). To complete the item, the test taker decides whether each of five shapes has been divided into fourths. The shapes vary. In order to answer correctly, the test taker must indicate shapes that meet the following two criteria: a) the overall shape has been subdivided into four subregions, and (b) the four subregions cover equal areas. Two of the shapes meet these two criteria, and three of the shapes do not. Initially, students' responses on each option were scored as a distinct, dichotomous variable according to whether it was correctly circled or not circled.

Because these five responses are part of a single item, we explored options for scoring the response dichotomously or polytomously. To score item 10 into one variable (thereby accounting for the item set), we considered scoring it as a single, polytomous variable by combining the scores of all the five responses additively to create a single score. Through a thought experiment considering the sample space of various outcomes, we decided this method had major flaws. As a reminder, a perfect score would result from a student selecting two of the items and not selecting three of the items. If student A circled one incorrect option but did not choose the other four options, student A would get two points out of the possible five. If student B chose one correct option and three incorrect options but did not choose the other correct option, student B would get one point out of the possible five. Although student A has a higher score than student B, student A's ability in fraction is not necessarily better than student B. If student C circled all of the five items, then student C would earn two out of five points, but it is not clear that student C demonstrated higher ability than student A (or vice versa). This thought experiment identified a major weakness in a potential decision to score this item with a single, polytomous variable.

Another method we considered was to score item 10 dichotomously as a single item. That is, a student would get one point only if he or she made correct choices on all five options in item 10 (i.e., selecting the two correct options, and not selecting the three incorrect options). Otherwise, the student would get zero point on item 10. Because the former scoring method was problematic in discriminating students' competence in fractions, we decided to score item 10 dichotomously in our analyses. Tables 1.1, 2.1, and 3.2 include information reflecting this item's change of scoring.

Table 3.2. Item Indexing and Scoring for both Test-Form and Final-Scale Format

Test-form item #	Scoring of test-form item	Final-scale item #	Scoring of final-scale item
1	0, 1, 2	1*	0, 1, 2
2	0, 1, 2	2*	0, 1, 2
3	0, 1	3*	0, 1
4	0, 1	4*	0, 1
5	0, 1	5*	0, 1
6	0, 1	6*	0, 1
7, 8, 9	0, 1	7*	0, 1, 2, 3
10	0, 1, 2, 3, 4, 5	8*	0, 1
11	0, 1, 2, 3, 4	9*	0, 1, 2, 3, 4
12	0, 1, 2, 3, 4	10*	0, 1, 2, 3, 4
13	0, 1	11*	0, 1
14	0, 1	12*	0, 1
15	0, 1	13*	0, 1
16	0, 1, 2, 3	14*	0, 1, 2, 3
17	0, 1	15*	0, 1
18	0, 1	16*	0, 1
19	0, 1	17*	0, 1
20	0, 1	18*	0, 1

Note. Test-form Item # = the item index from the original fraction test; Final-scale item # = the newly generated item number after combining items 7–9 (we differentiated test-form and final-scale item index by assigning a * after the final-scale item number).

4. Dimensionality Analysis

4.1. Exploratory Factor Analysis

To explore the dimensionality of the test, we first ran exploratory factor analysis (EFA) models using Mplus 7.0 (Muthén & Muthén, 1998–2012). Given that the dataset consisted of either dichotomously or polytomously scored variables, Mplus could run EFA based on an estimated polychoric correlation matrix. Adopting the Geomin rotation method and weighted least square estimation method with mean and variance adjusted (WLSMV; Finney & Distefano, 2013), Table 4.1 shows the eigenvalues estimated by Mplus with the corresponding percentages of explained variation. These eigenvalues are also illustrated in the scree plot in Figure 4.1. Based on these analyses, there appeared to be a single dominant factor in the data.

Table 4.1. Eigenvalues Estimated from Mplus and Their Corresponding Percentages of Explained Variation

Component	Eigenvalue	% Variation explained
1	8.64	48.00
2	1.15	6.39
3	0.93	5.17
4	0.91	5.06
5	0.78	4.33
6	0.68	3.78
7	0.66	3.67
8	0.59	3.28
9	0.52	2.89
10	0.51	2.83
11	0.44	2.44
12	0.42	2.33
13	0.40	2.22
14	0.34	1.89
15	0.32	1.78
16	0.28	1.56
17	0.25	1.39
18	0.16	0.89

Note. Component = the component index; Eigenvalue = the eigenvalue associated with a given component estimated by Mplus; % Variation Explained = the percentage of variation explained by a given component in the data.

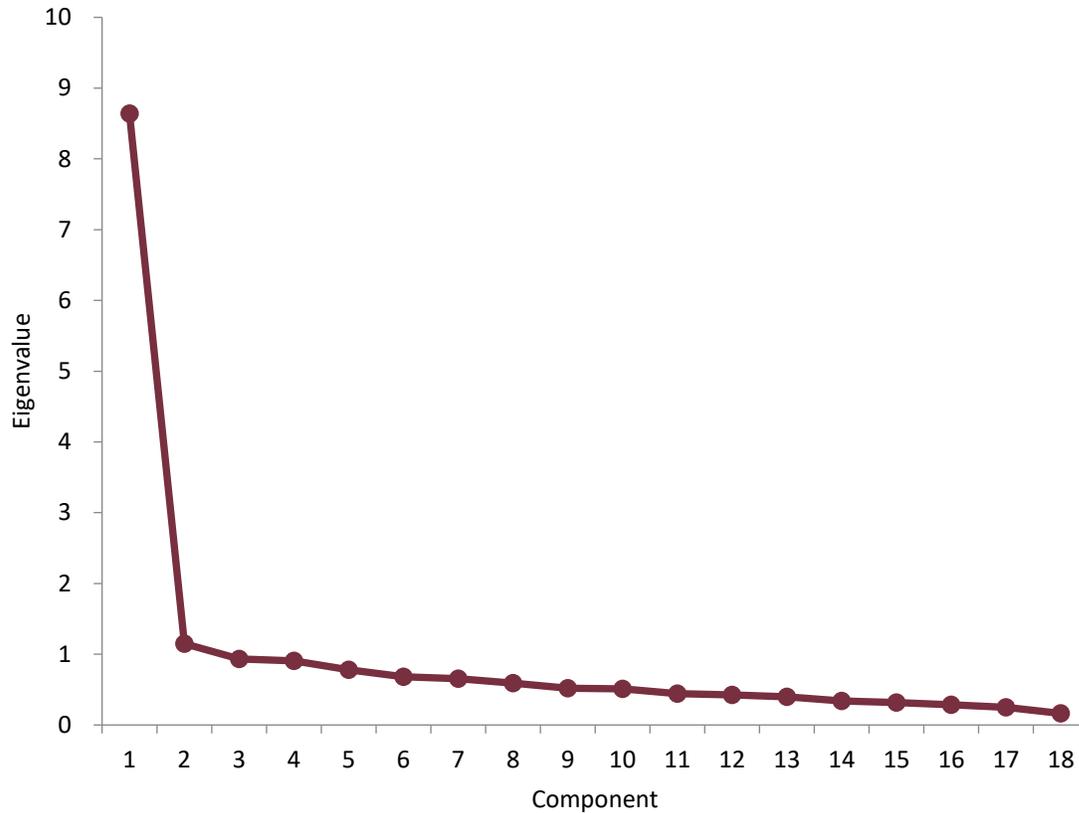


Figure 4.1. Scree plot of eigenvalues estimated from Mplus.

4.2. Parallel Analysis

To further evaluate dimensionality, we performed parallel analysis. Parallel analysis (PA) is a procedure to yield optimal solutions to the number of components problem in EFA, and it is considered superior to rule-of-thumb procedures (Wood, Tataryn, & Gorsuch, 1996; Zwick & Velicer, 1982, 1986) such as Kaiser’s rule (Kaiser, 1960). The idea of PA is to select those components that account for more variance than those generated from random data (O’Connor, 2000). We used the *psych* (Revelle, 2017) package in R 3.4.0 (R Core Team, 2017) to perform PA. The PA involved the use of principal component analysis with 200 simulated analyses on polychoric correlation matrices. The results of the PA were consistent with the previous EFA results in that the data appeared to be unidimensional. The convergence and lack of ambiguity in these results appeared to support an assumption of unidimensionality in the data.

5. Classical Testing Theory (CTT) Analyses

Based on the results described in the previous section, we adopted a unidimensional data structure. The next step in our analysis was to analyze the test using a framework based on classical testing theory (CTT) using SPSS 22.0 (IBM corp., 2013).

5.1. Distribution of the Observed Test Score

We first examined the characteristics of the total test score. Results indicated that the mean of the total test score was 11.86 with a standard deviation of 7.09. In addition, both the median and the mode of the test score were 11.00. Figure 5.1 displays the distribution of the observed test scores in the final-scale format. Note that, although the final-scale format has only 18 items, the observed test scores ranged from 0 to 30, because the format had several item sets (i.e., items 1*, 2*, 7*, 9*, 10*, 14*) that were ultimately collapsed into polytomous items. Table 2.1 has information showing the detail for the scoring of each item in the final-scale format.

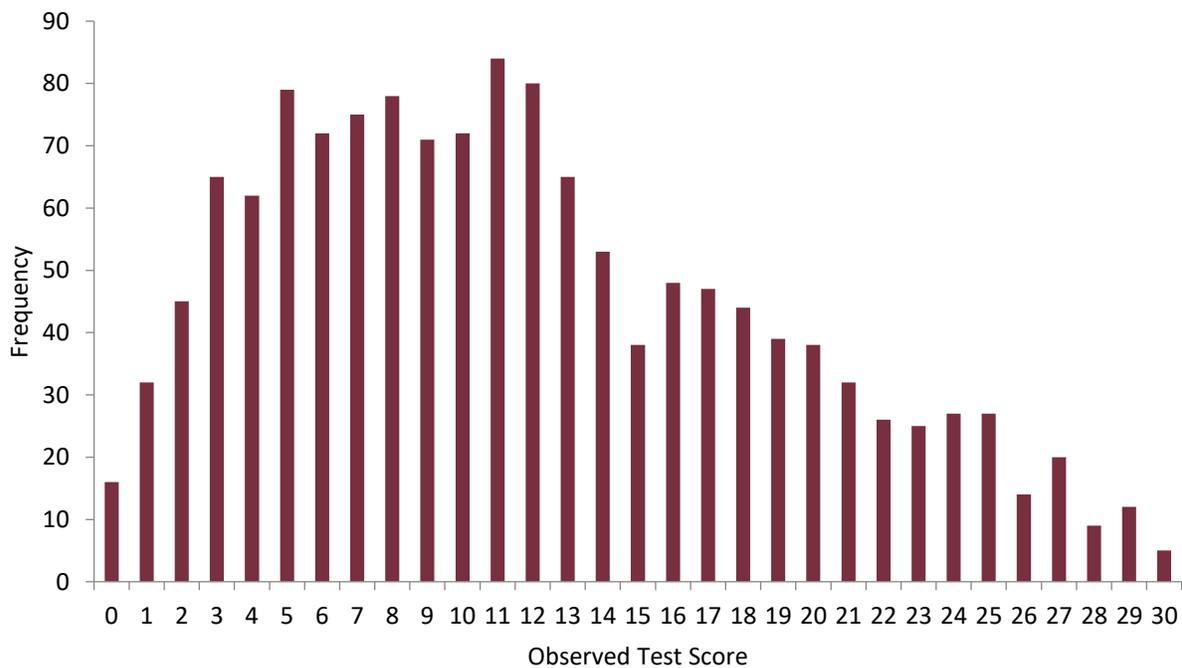


Figure 5.1. Bar graph depicting the distribution of the observed test score in the final-scale format.

5.2. Item Difficulty & Discrimination

Next, we calculated the item difficulty and item discrimination for each of the final-scale items using a CTT-based approach. For both dichotomous and polytomous items, the values of p could be calculated using the formula in Equation 1 (McDonald, 1999). When the items were dichotomously coded, the values of p were simplified to the proportion of correct answers,

$$p = \frac{\text{ItemMean} - \text{ItemMin}}{\text{Theoretical Score Range}} \quad (1)$$

where p is the symbol of the item difficulty index.

The item difficulty indices varied from a minimum of .12 (item 15*) to a maximum of .81 (item 4*). To investigate item discrimination, we calculated the item-rest correlation coefficients (i.e., corrected item-total correlation coefficients) for each of the items. Item-rest correlation is defined as the Pearson product-moment correlation between the score of the focal item and the test score, which excludes the score of the focal item (MacDonald, 1999). For dichotomous items, their item-rest correlations are point-biserial correlations. For polytomous items, their item-rest correlations are point-polyserial correlations. All the items (except for item 13*) had item-rest correlation coefficients larger than .38, suggesting that the items generally had good discriminative power. Table 5.1 shows the results of calculations of item difficulty and item discrimination as well as some descriptive information for each of the test items.

Table 5.1. Item Difficulty and Discrimination from CTT Analyses

Final-scale item #	<i>M</i>	<i>SD</i>	<i>p</i>	Item-Rest <i>r</i>
1*	1.20	0.89	.60	.45
2*	0.74	0.85	.37	.60
3*	0.73	0.45	.73	.47
4*	0.81	0.39	.81	.39
5*	0.53	0.50	.53	.48
6*	0.35	0.48	.35	.58
7*	0.85	1.29	.28	.54
8*	0.34	0.47	.34	.43
9*	2.55	1.50	.64	.58
10*	1.50	1.37	.38	.69
11*	0.61	0.49	.61	.54
12*	0.13	0.33	.13	.47
13*	0.23	0.42	.23	.29
14*	0.54	0.83	.18	.52
15*	0.12	0.33	.12	.38
16*	0.20	0.40	.20	.42
17*	0.15	0.36	.15	.38
18*	0.27	0.44	.27	.57

Note. Final-scale item # = the newly generated item number after combining items 7–9 (we differentiated test-form and final-scale item index by assigning a * after the final-scale item number); p = item difficulty; Item-Rest r = item-rest correlation coefficient (i.e., corrected item-total correlation coefficient), which is the Pearson correlation between the item score and the test score that excludes the item score.

5.3. Reliability & Standard Error of Measurement

Because evidence supported the unidimensionality assumption, we chose Coefficient α (Cronbach, 1951) to estimate the reliability of the test. Coefficient α is the average of all the possible split half reliabilities of test data, correcting for test length. According to the SPSS outputs, the Coefficient α for the data in the present sample using the final-scale format was .85. We subsequently calculated the standard error of measurement (SEM) for the present test data in the final-scale format. SPSS output indicated that the scale variance was 50.23. Using the formula presented in Equation 2, SEM was calculated to be 2.74, where σ^2 is the test variance, and ρ_{XX} is the Coefficient α of the test.

$$SEM = \sqrt{\sigma^2 \times (1 - \rho_{XX})}, \quad (2)$$

6. Item Response Theory (IRT) Analyses

6.1. Model Description

We used flexMIRT 3.5 (Cai, 2017) to perform the following IRT analyses. For multiple-choice items (i.e. items 3*, 4*, and 6*) that were scored dichotomously, a three-parameter (3PL) model was used to fit the data. We chose 3PL for those items, because they are 4-option multiple-choice items, where student guessing should be of concern. For the other dichotomously scored items (i.e. items 5*, 8*, 11*, 12*, 13*, 15*, 16*, 17*, and 18*), a two-parameter (2PL) model was used. We chose 2PL for these items given that all of these items were constructed-response items (where guessing should not be of concern), with the exception of item 8*. Item 8* consists of five dichotomous responses (i.e., circled, not circled). Given the small possibility of the guessing success (i.e., 0.5^5), we decided it was not necessary to model the guessing parameter for this item. For the polytomously scored items (i.e. item 1*, 2*, 7*, 9*, 10*, and 14*), a Generalized Partial Credit Model (GPCM) was used.

Results of FlexMIRT indicated that successful convergence was reached in the computation, and the value of $-2\log\text{likelihood}$ was 31738.69. The formulas of the 2PL model, 3PL model, and GPCM are shown and explained subsequently based on the parameterization of de Ayala (2009).

The formula of the 2PL model is presented in Equation 3,

$$P_j(\theta) = \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]} \quad (3)$$

where

a_j is the discrimination index of item j ($j = 1, 2, \dots, J$),

b_j is the difficulty index of item j ,

P_j is the probability of correct answer,

θ is the person ability.

The formula of the 3PL model is presented in Equation 4,

$$P_j(\theta) = g_j + (1 - g_j) \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]} \quad (4)$$

where

a_j is the discrimination index of item j ($j = 1, 2, \dots, J$),

b_j is the difficulty index of item j ,

P_j is the probability of correct answer,

θ is the person ability,

g_j is the guessing parameter of item j .

The formula of the GPCM is presented in Equation 5,

$$P_{jk}(\theta) = \frac{\exp \sum_{h=0}^k [a_j(\theta - b_j + d_{jh})]}{\sum_{c=0}^{m_j} \exp \sum_{h=0}^c [a_j(\theta - b_j + d_{jh})]}, \quad (5)$$

where

a_j is the discrimination index of item j ($j = 1, 2, \dots, J$),

b_j is the overall difficulty index of item j ,

P_{jk} is the probability of correct answer,

θ is the person ability,

d_{jh} is deviation from overall item difficulty b_j , i.e., distance from overall item difficulty to the h^{th} threshold, k is item category, $k \in \{0, 1, 2, \dots, m_j\}$.

6.2. Item Difficulty and Discrimination

Table 6.1 presents the results regarding the distribution of both item difficulty and item discrimination estimated from the final-scale format. The item discrimination estimate ranged from 0.65 to 3.56. The item difficulty index ranged from -0.88 to 1.89 . Tables 6.2, 6.3, and 6.4 present parameter estimates for each item based on the 2PL, 3PL, or GPCM models, respectively. Figure 6.1 displays the item discrimination estimates of all the items. The discrimination indices for all the 18 items were greater than 0.50, and 13 of the items had values above 1.00 (i.e., items 2*, 3*, 4*, 5*, 6*, 8*, 10*, 11*, 12*, 15*, 16*, 17*, and 18*). The highest discrimination value was from item 6*. The estimated discrimination index of item 6* was 3.65 with a standard error of 0.46. Figure 6.2 displays the item difficulty estimates of all the items. Six items (i.e. items 1*, 3*, 4*, 5*, 9*, and 11*) had b values below 0.00, and 12 items (i.e. items 2*, 6*, 7*, 8*, 10*, 12*, 13*, 14*, 15*, 16*, 17*, and 18*) had b values above 0.00.

Table 6.1. Descriptive Statistics of Discrimination Index and Difficulty Index of all the 18 Items

	<i>M</i>	<i>SD</i>	Min	Max	Skewness	Kurtosis
<i>a</i>	1.50	0.74	0.65	3.65	1.50	2.95
<i>b</i>	0.59	0.90	-0.88	1.89	-0.11	-1.33

Note. a = item discrimination index; b = item difficulty index.

Table 6.2. Parameter Estimates and Standard Errors for Final-Scale Items Modeled Using 2PL

Final-scale item #	a (SE)	b (SE)
5*	1.45 (0.10)	-0.13 (0.05)
8*	1.19 (0.09)	0.69 (0.07)
11*	2.01 (0.15)	-0.39 (0.05)
12*	1.91 (0.14)	1.55 (0.08)
13*	0.78 (0.08)	1.78 (0.18)
15*	1.39 (0.13)	1.89 (0.13)
16*	1.31 (0.11)	1.37 (0.10)
17*	1.27 (0.11)	1.71 (0.12)
18*	2.08 (0.15)	0.79 (0.05)

Note. Final-Scale Item # = the newly generated item number after combining items 7–9 (we differentiated test-form and final-scale item index by assigning a * after the final-scale item number); a = item discrimination index; b = item difficulty index; SE = standard error.

Table 6.3. Parameter Estimates and Standard Errors for Final-Scale Items Modeled using 3PL

Final-scale item #	a (SE)	b (SE)	g (SE)
3*	2.29 (0.25)	-0.57 (0.10)	0.17 (0.05)
4*	2.14 (0.26)	-0.88 (0.14)	0.25 (0.08)
6*	3.65 (0.46)	0.61 (0.05)	0.10 (0.02)

Note. Final-Scale Item # = the newly generated item number after combining items 7–9 (we differentiated test-form and final-scale item index by assigning a * after the final-scale item number); a = item discrimination index; b = item difficulty index; g = item guessing parameter; SE = standard error.

Table 6.4. Parameter Estimates and Standard Errors for Final-Scale Items Modeled Using GPCM

Final-scale item #	a (SE)	b (SE)	d_1 (SE)	d_2 (SE)	d_3 (SE)	d_4 (SE)
1*	0.77 (0.06)	-0.44 (0.06)	-0.87 (0.13)	0.87 (0.13)		
2*	1.17 (0.08)	0.47 (0.05)	-0.12 (0.07)	0.12 (0.07)		
7*	0.65 (0.04)	0.83 (0.06)	-4.32 (0.44)	2.20 (0.36)	2.12 (0.26)	
9*	0.91 (0.06)	-0.49 (0.04)	-0.28 (0.13)	0.39 (0.13)	0.12 (0.11)	-0.23 (0.09)
10*	1.08 (0.07)	0.42 (0.04)	0.97 (0.07)	-0.11 (0.08)	-0.58 (0.10)	-0.28 (0.11)
14*	0.99 (0.08)	1.39 (0.07)	0.66 (0.07)	-2.11 (0.26)	1.45 (0.27)	

Note. Final-Scale Item # = the newly generated item number after combining items 7–9 (we differentiated test-form and final-scale item index by assigning a * after the final-scale item number); a = item discrimination index; b = item difficulty index; d_h ($h = 1, 2, 3, 4$) = deviation from the overall item difficulty; SE = standard error.

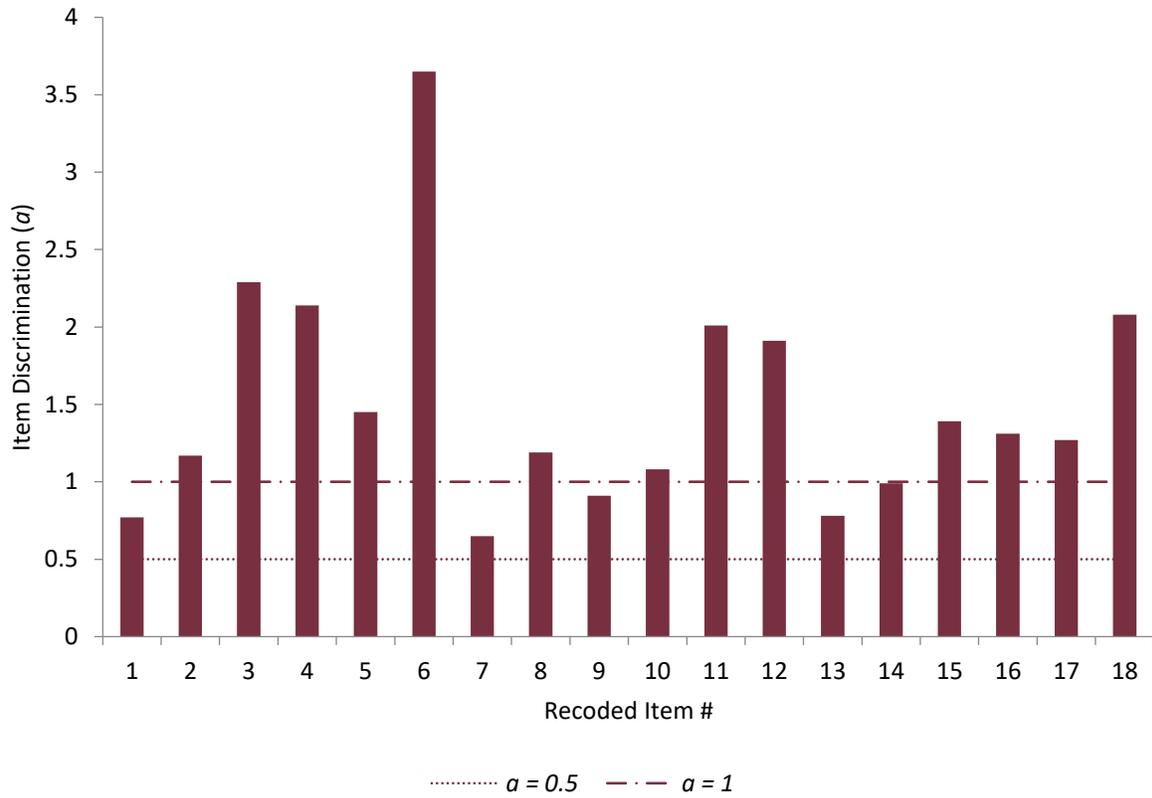


Figure 6.1. Item discrimination estimate (a) of each final-scale item.

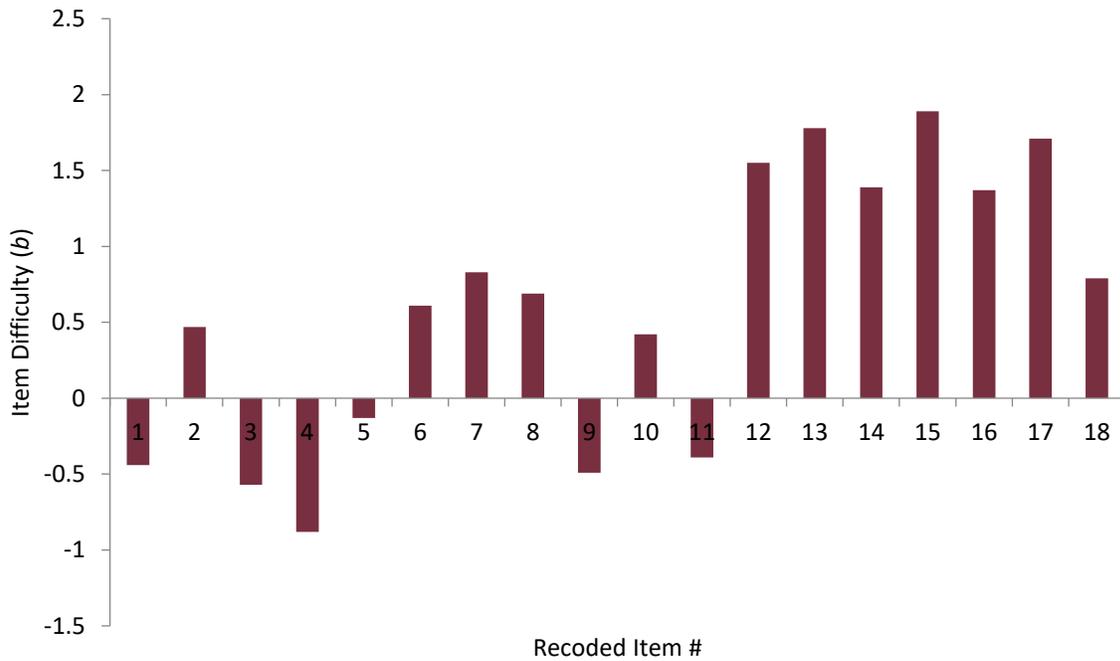


Figure 6.2. Item difficulty estimate (b) of each final-scale item.

6.3. Test Information and Estimated Person Ability

Figure 6.3 displays the resulting test information curve and the CSEM for the test in the final-scale format. The formula used for the calculation of CSEM were in accordance with recommendations made by de Ayala (2009). Equation 6 shows the formula used in the CSEM calculation, where I is the test information function for a given person ability, and θ is the person ability.

$$CSEM(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{6}$$

Given the relationship between test information and CSEM, a person ability (i.e., θ) estimate around the value of 0.80 was associated with the largest test information and the lowest CSEM. In addition, the CSEM curve in Figure 6.3 suggests that the person ability estimate was related to the lowest CSEM (i.e., highest accuracy of person ability estimation) when it ranged between -0.80 and 1.60 , but it was related to the highest CSEM (i.e., lowest accuracy of person ability estimation) when it was less than -2.00 .

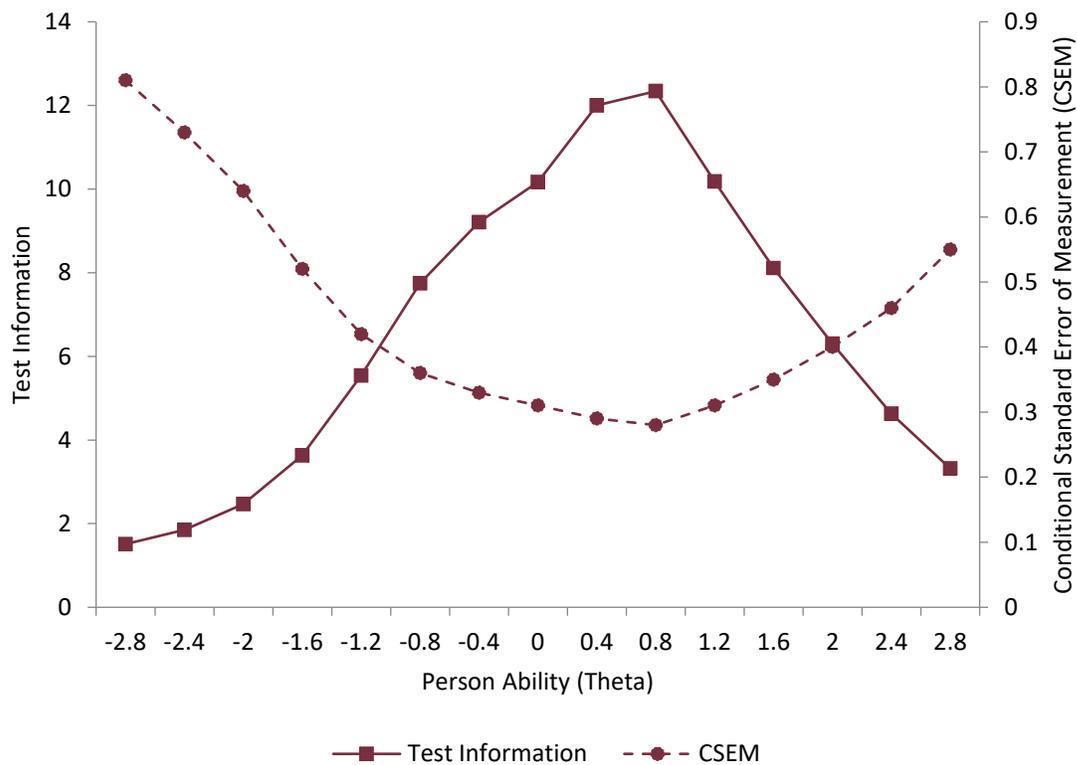


Figure 6.3. Test information curve and conditional standard errors of measurement (CSEM) for the final-scale format.

We used maximum likelihood estimation (MLE) to estimate the latent person ability of each student. Figure 6.4 illustrates the distribution of person ability using MLE. The mean and standard deviation were -0.07 and 1.36 , respectively. The skewness and kurtosis estimates were -0.92 and 5.08 , respectively.

The spikes at the higher and lower end of the horizontal axis of the distribution curve were a result of some students having perfect scores or zero scores (whose MLE estimates were not available), respectively. Sixteen of the 1,400 students did not respond correctly to any of the items. Ten of these students were in grade 3, while six of these students were in grade 4. Five of the 1,400 students responded correctly to every item (i.e., perfect score). Two of these students were in third grade, and three of them were in fourth grade.

We also used *expected a posteriori* (EAP) method to estimate the person ability of each student. Figure 6.5 illustrates the distribution of person ability using EAP. The sample distribution of person ability scores ranged from -1.95 to 2.67 . The mean and standard deviation were 0.00 and 0.94 , respectively. The skewness and the kurtosis estimates were 0.31 and -0.25 , respectively. The distribution of person ability estimates using EAP does not appear to be that much different from the standard normal distribution in terms of mean, standard deviation, skewness and kurtosis.

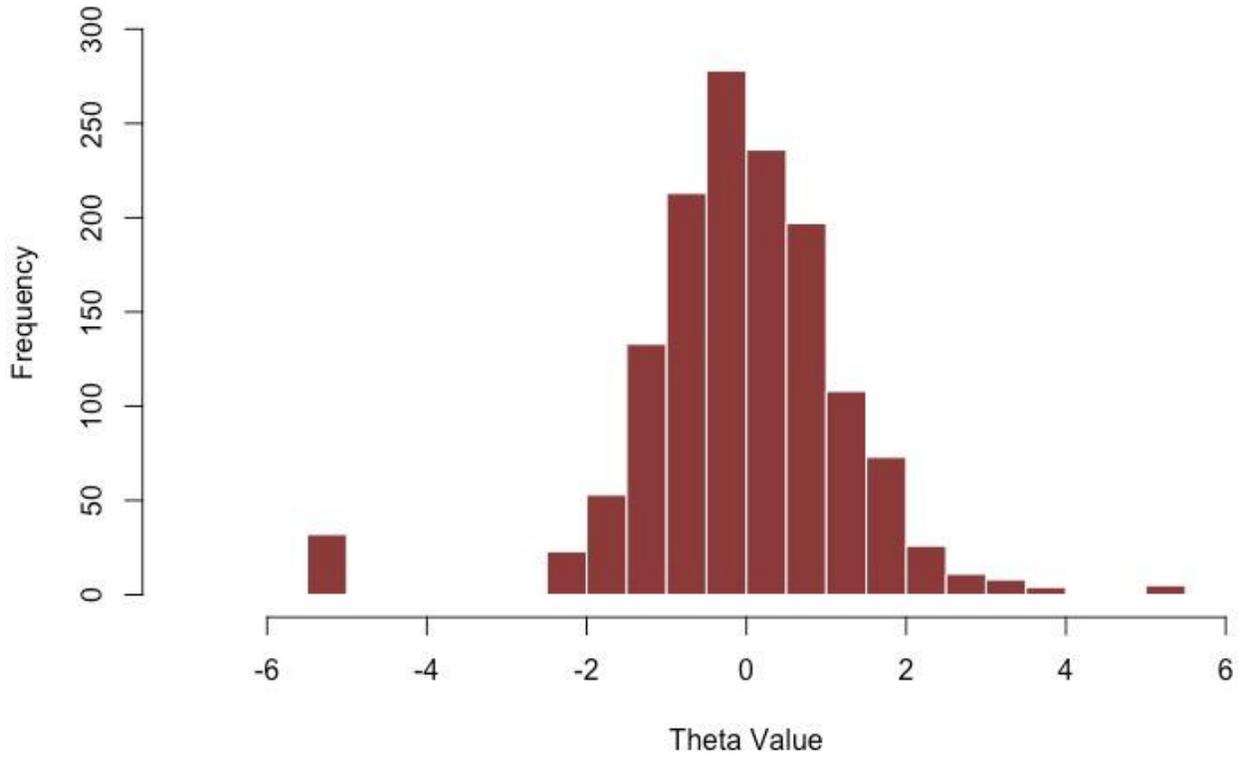


Figure 6.4. Person abilities (i.e., ϑ) estimated by maximum likelihood estimation (MLE).

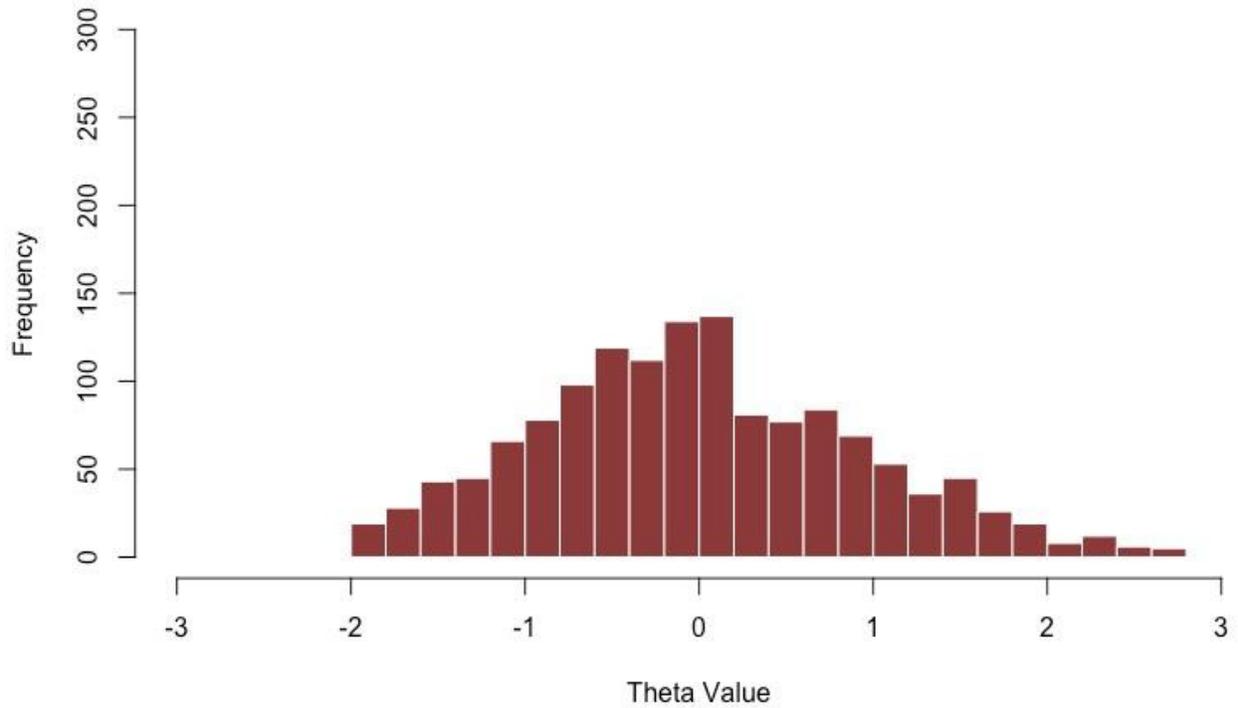


Figure 6.5. Person abilities (i.e., ϑ) estimated by expected a posteriori (EAP).

7. Additional Analyses

7.1. Intraclass Correlation Coefficient

The intended use of the scores from the Early Fractions Test were to serve as a student achievement pretest for a large-scale, randomized-controlled trial. Empirical estimates of intraclass correlation coefficients (ICC) can provide important guidance and insight during the power analysis phase in the design of similar studies. For that reason, we calculated the ICC and report it here.

To calculate the ICC, we used HLM7 software and specified a two-level, unconditional model with students at level 1, schools at level 2, and student test score as the dependent variable. We divided the between-school variance by the total variance (i.e., between-school plus within-school) to arrive at the ICC estimate. We calculated the ICC three times for three different estimates of student ability: the total raw score, the person ability (theta) estimate using the ML method, and the person ability (theta) estimate using the EAP method.

With the sample of 1,400 students representing 66 schools, the school-level ICC estimate based on the total raw scores in this sample was .44. The ICC estimate based on the person ability (theta) estimates generated by the ML estimator was .37. The ICC estimate based on the person ability (theta) estimates generated by the EAP estimator was .45.

7.2. Predictive Validity

The ability estimates generated in the fall 2016 administration of the Early Fractions Test are designed to be used in a larger study involving a randomized controlled trial. They will be used to test for baseline equivalence of the schools assigned to the treatment conditions and as a student achievement pretest covariate in multilevel models of analysis of covariance. Having the student posttest scores available to us, we calculated how much of the variance in student posttest scores (administered in spring 2017) was explained by those same students' scores on the fall 2016 Early Fractions Test. Like the ICC estimates, this information can be useful in the power analysis phase of research design. It also can provide some evidence of external validity (Flake, Pek, & Hehmann, 2017).

Based on a sample of 1,134 students who completed both the pretest and the posttest, and using SPSS version 22, we found a Pearson correlation coefficient of .66 between the total raw score on the pretest and the total raw score on the posttest. The coefficient was .69 for the theta scores based on the EAP estimator and .61 for the theta scores based on the ML estimator. Therefore, with no adjustment for other factors such as clustering in schools, the student ability estimates from the Early Fractions Test used at the beginning of the school year explains somewhere between 37% and 48% of the variance in student scores measured at the end of the school year for these grade 3 or 4 students representing 66 schools. We note that the pretest and posttest scores used in these analyses were not equated.

8. Discussion and Conclusions

The present report addresses components of validity corresponding to substantive, structural, and external validity (Flake, Pek, & Hehman, 2017; Loevinger, 1957; Benson, 1998). Focused on scoring and data modeling, the majority of the present report focused on the structural component. Chapter 2 briefly discusses the content of the test. Because the purpose of the test is to serve as a student pretest covariate in models estimating the effect of an intervention on student posttest scores, we also began to examine the extent to which scores on this test explain the variance in student scores on the posttest (a matter of external validity). Other elements of an external validity argument will be explored when the test scores are used in attempt to detect differences among students in different treatment conditions, which a previous version of this test appears to have been well-suited to do (Lewis & Perry, 2017).

Several of the responses in the Early Fractions Test were presented as testlets (i.e., item sets), which introduces a potential threat to the validity of an assumption of local-independence. We found evidence of collinearity of items in item sets when the items were considered as separate items scored dichotomously. We also found that collapsing those responses into polytomous variables preserved the assumption of independence among items and resulted in items with good parameters.

Results of Exploratory Factor Analysis and Parallel Analysis both support the assertion that the Early Fractions Test is measuring a single, dominant trait. Coding the item sets as polytomous variables resolved problems of collinearity among items and improved the ability to interpret item-level scores. Taken together, we find consistent evidence supporting an assumption of unidimensionality in the data in the 18-item test.

We analyzed the dataset following both CTT approach and IRT approach. For the CTT analyses, they reflect typical procedures performed at both test and item levels given the unidimensionality assumption in the data (MacDonald, 1999). Our IRT analyses include a series of decisions guided by both empirical evidence based on the sample data and *a priori* recommendations (de Ayala, 2009). Specifically, we chose different IRT models based on types of response prompted by the item format, and we adjusted item coding when such a procedure was empirically and computationally necessary. Lastly, to score student ability using IRT, we adopted MLE as the main estimator but also used EAP as a supplementary estimator when MLE failed in specific cases.

Findings from the CTT and IRT analyses indicate that the test items were moderately difficult for the student sample and had good discriminative power. According to the CTT results, the discrimination index estimates for all the items were above .29, and the estimate of item difficulty index ranged from .12 to .81. Eight items had estimates of difficulty index less than .30, which can be interpreted as difficult items (Kubiszyn & Borich, 2013). According to the IRT results, all the items had discrimination index estimates above 0.50, and 13 of the 18 items had discrimination index estimates above 1.00. Regarding the estimates of item difficulty index, 12 out of 18 items had a positive value for the estimated difficulty levels, and 6 items had difficulty estimates less than 0.00.

Approximately 1% of the student sample did not provide a correct response to any of the items. These students were balanced proportionally across both the third and fourth-grade subsamples. This suggests that a future version of the test might want to include items at a lower difficulty level to discriminate among students in the lowest percentile of the ability distribution. A few students also received perfect scores, suggesting that a few high-difficulty items might also be warranted.

Altogether, our evaluation of the structural validity of the resulting 18-item scale supports the assertion that the Early Fractions Test meets or exceeds the usual standards for educational research for its stated purpose.

References

- Beckmann, S. (2005). *Mathematics for elementary teachers*. Boston, MA: Pearson Education.
- Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, 17, 10–17.
- Cai, L. (2017). flexMIRT R version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- California Department of Education. (n.d.). CST released test questions: Released test questions for the California Standards Tests (CSTs). Retrieved from <http://www.cde.ca.gov/ta/tg/sr/css05rtq.asp>
- Cronbach, L. J. (1951). *Coefficient alpha and the internal structure of tests*. *Psychometrika*, 16, 297–334.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.
- IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.
- Finney, S. J. & Distefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In Hancock & Mueller (Ed.), *Structural equation modeling a second course* (pp. 439-492). Charlotte, NC: Information Age Publishing, Inc.
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 1–9.
- 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.
- Hironaka, H., & Sugiyama, Y. (2006). *Mathematics for elementary school, Grades 1–6*. Tokyo, Japan: Tokyo Shoseki.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141–151.
- Kubiszyn, T., & Borich, G. (2015). *Educational testing and measurement (10th ed.)*. Hoboken, NJ: John Wiley & Sons.
- Lewis, C. 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.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Muthén, L. K. & Muthén, B. O. (1998–2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods*, 32(3), 396–402.

- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Revelle, W. (2017) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, <https://CRAN.R-project.org/package=psych> Version = 1.7.5.
- Schoen, R. C., Anderson, D., Riddell, C. M., & Bauduin, C. (2017). Elementary Mathematics Student Assessment: Measuring the performance of grade 4, 5, and 6 students in problem solving and computation involving whole number, fractions, and equality in fall 2015 (Research Report No. 2017-21). Tallahassee, FL: Learning Systems Institute, Florida State University.
- Schoen, R. C., Anderson, D., Riddell, C. M., & Bauduin, C. (2017). Elementary Mathematics Student Assessment: Measuring the performance of grade 4, 5, and 6 students in problem solving and computation involving whole number, fractions, and equality in spring 2016 (Research Report No. 2017-23). Tallahassee, FL: Learning Systems Institute, Florida State University.
- Van de Walle, J. A. (2007). Developing fraction concepts. In *Elementary and middle school mathematics: Teaching developmentally* (6th ed., pp. 293–315). Boston, MA: Pearson Education.
- Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, 1(4), 354–365.
- Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, 17(2), 253–269.
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3), 432–442.

Appendix A. The Early Fractions Test Form

Student Fractions Questions

2016-2017

Student Name: _____

Teacher Name: _____

School: _____ **Grade level:** _____ **Date:** _____

[Page intentionally left blank]

This paper may include some kinds of problems that are new or hard for you. Don't worry if you can't solve them. You won't be graded on this test, but the test will help us understand our math program.

Please try your hardest!

1)



2)



Answer: _____ 



3

3)  Answer: _____



4) 

Answer: _____



5)  Answer: _____

6)  Answer: _____



Write your answers to the following problems:

7)  Answer: _____

8)  Answer: _____

9)  Answer: _____

10) 


5

11)

[Redacted]

[Redacted]

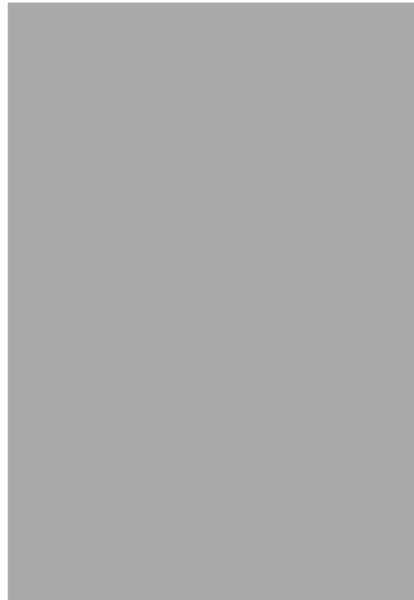
[Redacted]

[Redacted]

Answer: _____ [Redacted]

[Redacted]

12)



13)



Answer: _____

14)



Answer: _____

7

15)



16)



8

17)



9

18)



19)



20)



10

Appendix B. Administration Instructions

Instructions for Administration of the “Early Fractions Test” Posttest

Overview

Thank you for your participation in the study *Improvement of Elementary Fractions Instruction*. This document provides instructions for giving the “Early Fractions Test” posttest. Please give this test to your class at your earliest convenience. A pre-paid mailing label is included for returning the posttest to us. Please do not hesitate to contact Claire Riddell (criddell@lsi.fsu.edu) if you have any questions about any aspect of the posttest.

Materials Needed for Testing

The following materials are needed for the posttest:

- One copy of the “Early Fractions Test” posttest for each student (provided)
- At least one sharpened pencil for each student

Testing

The “Early Fractions Test” posttest is designed to be given to your whole class at once, with students completing the test independently. Students write their answers directly on the test. Give the posttest as you would other student tests—for example, have students space out desks or use student “privacy folders” if that is what they usually do.

Please administer the posttest according to the following guidelines:

- Check that all students fill out the information box on the cover page.
- Let students know that no talking or communication between students is permitted during testing.
- Read students just the information at the top of the posttest:
This paper may include some kinds of problems that are new or hard for you. Don’t worry if you can’t solve them. You won’t be graded on this test, but the test will help us understand our math program. Please try your hardest!
- 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.

To ensure validity of the posttest, we also ask that you keep the tests private, in a secure location, before testing and until they are returned to us.

Accommodations

Students with special academic plans (e.g., IEP, 504, ELL) may receive the appropriate testing accommodations as specified in their plans.

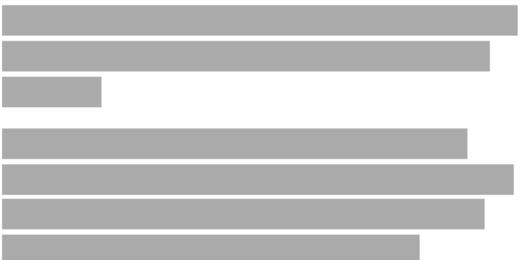
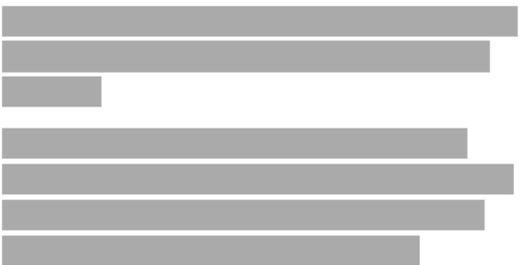
Testing Time Allocation

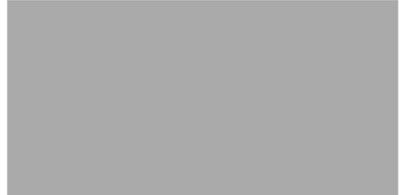
This is not intended to be a timed test, and students should be allowed adequate time to answer the questions. We anticipate that administration of this posttest will require approximately 30-40 minutes.

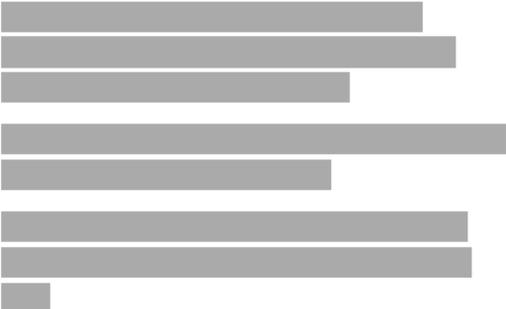
Submitting the Early Fractions Test Materials

Upon conclusion of testing, place all test booklets (both used and unused) in the box you received the materials in along with your completed Class Roster. Place the pre-paid mailing label on the box and drop it off at a UPS store location, or “Schedule a Pickup” with UPS at www.ups.com.

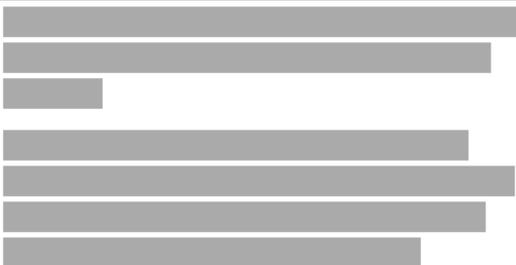
Appendix C. Scoring Key

Item	Data Entry	Scoring Criteria
<p>Item 1a: Rabbit (Numeric Answer)</p> 	<p>Enter the response as written</p>	
<p>Item 1b: Rabbit (Number Line)</p> 	<p>Score using overlay and scoring criteria): Enter 1 for correct, 0 for incorrect, DNS, or UI</p>	
<p>Item 2a: Polar Bear (Numeric Answer)</p> 	<p>Enter the response as written</p>	
<p>Item 2b: Polar Bear (Number Line)</p> 	<p>Score using overlay and scoring criteria): Enter 1 for correct, 0 for incorrect, DNS, or UI</p>	

<p>Item 3: Part of Referent Unit (2/3)</p> 	<p>Enter letter corresponding with student's response, DNS, or UI</p>	
<p>Item 4: Partitioned Bars (1/6)</p> 	<p>Enter letter corresponding with student's response, DNS, or UI</p>	
<p>Item 5: 1 gallon vs. 5/6 gallon</p>  Answer: _____	<p>Enter the response as written</p>	
<p>Item 6: Determine the Greatest Fraction (MC)</p> 	<p>Enter letter corresponding with student's response, DNS, or UI</p>	
<p>Item 7:   Answer: _____</p>	<p>Enter the response as written</p>	
<p>Item 8:   Answer: _____</p>	<p>Enter the response as written</p>	

<p>Item 9:   Answer: _____</p>	<p>Enter the response as written</p>	
<p>Item 10: Equal Partitioning (Fourths) </p>	<p>Enter 1 for circled or 0 for not circled for each shape. UI is a valid data code for this item,</p>	
<p>Item 11a: 1/3 of the Shaded Ribbon </p>	<p>Enter the response as written</p>	
<p>Item 11b: Shade 1/2 </p>	<p>Score using overlay and scoring criteria): Enter 1 for correct, 0 for incorrect, DNS, or UI</p>	

<p>Item 11c: Shade $\frac{3}{4}$ </p>	<p>Score using overlay and scoring criteria): Enter 1 for correct, 0 for incorrect, DNS, or UI</p>	
<p>Item 11d: Shade $\frac{5}{6}$ </p>	<p>Score using overlay and scoring criteria): Enter 1 for correct, 0 for incorrect, DNS, or UI</p>	
<p>Item 12a: Three $\frac{1}{4}$ pieces </p>	<p>Enter the response as written</p>	
<p>Item 12b: Three of what in $\frac{3}{5}$ </p>	<p>Enter the response as written</p>	

<p>Item 12c: How many $\frac{1}{10}$ in $\frac{7}{10}$</p> 	<p>Enter the response as written</p>	
<p>Item 12d: Blank Over $8 = 1$</p> 	<p>Enter the response as written</p>	
<p>Item 13: How Many Fourths Make a Whole</p> 	<p>Enter the response as written</p>	
<p>Item 14: How Many $\frac{1}{4}$ cups are in 3 cups?</p> 	<p>Enter the response as written</p>	
<p>Item 15: $\frac{3}{4}$ on a Number Line</p> 	<p>Score using overlay and scoring criteria): Enter 1 for correct, 0 for incorrect, DNS, or UI</p>	

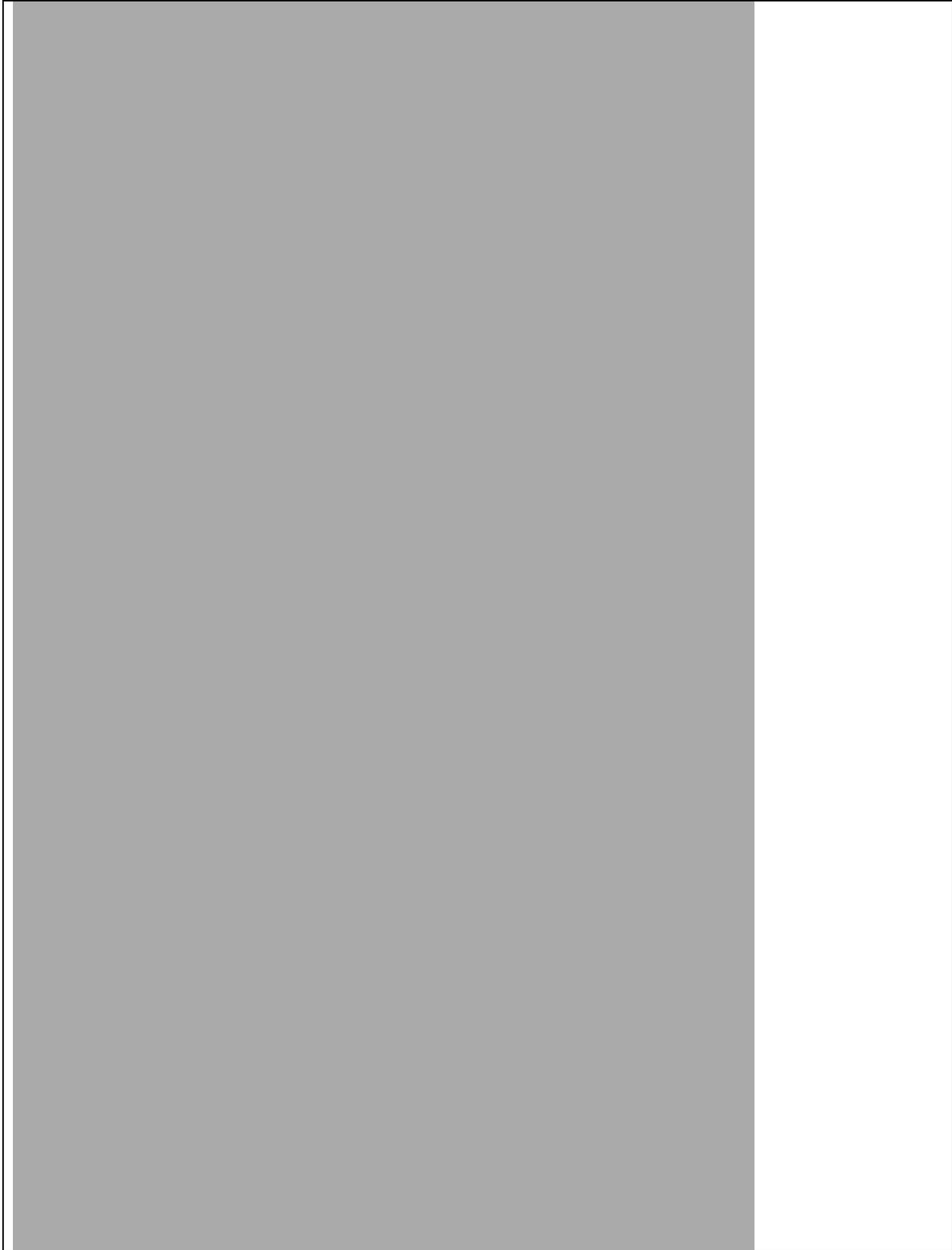
<p>Item 16A: Determine $9/8$ on the Number Line</p> 	<p>Enter the response as written</p>	
<p>Item 16B: Determine 2 on the Number Line</p> 	<p>Enter the response as written</p>	
<p>Item 16C: Determine $22/8$ on the Number Line</p> 	<p>Enter the response as written</p>	
<p>Item 17: Jose and Ella's Pizzas</p> 	<p>Score using NAEP criteria and enter the applicable score (1–5). DNS is a valid data code for this item.</p>	
<p>Item 18: Determining Referent Unit from $3/5$ Bar</p> 	<p>Score using overlay and scoring criteria): Enter 1 for correct, 0 for incorrect, DNS, or UI)</p>	

<p>Item 19: Draw a Bar that is $\frac{4}{3}$</p> 	<p>Score using overlay and scoring criteria): Enter 1 for correct, 0 for incorrect, DNS, or UI)</p>	
<p>Item 20: Joe's Walk</p> 	<p>Enter 1 for correct, 0 for incorrect, DNS, or UI</p>	

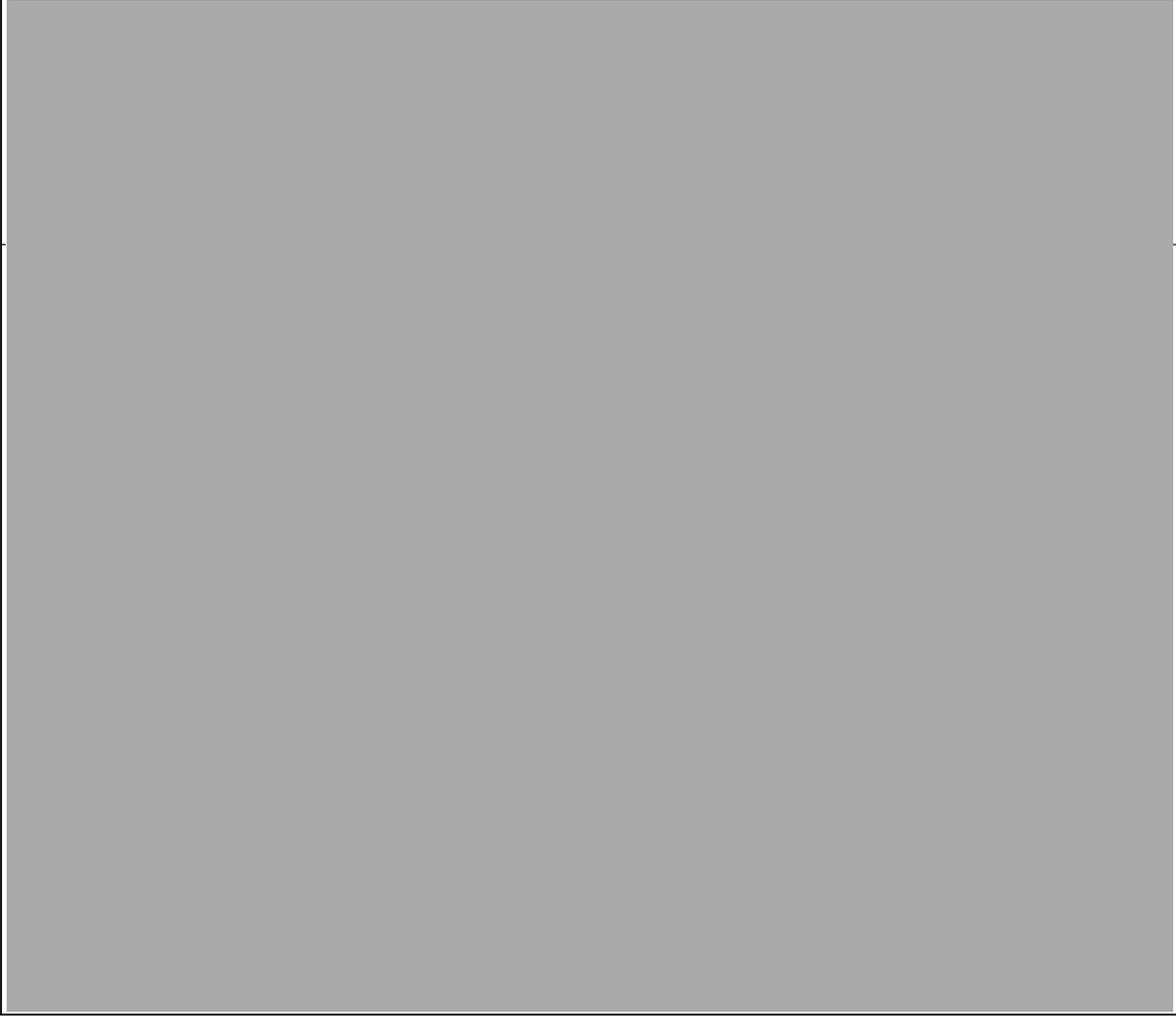
Item 17 (Jose and Ella's Pizzas) Scoring Rubric

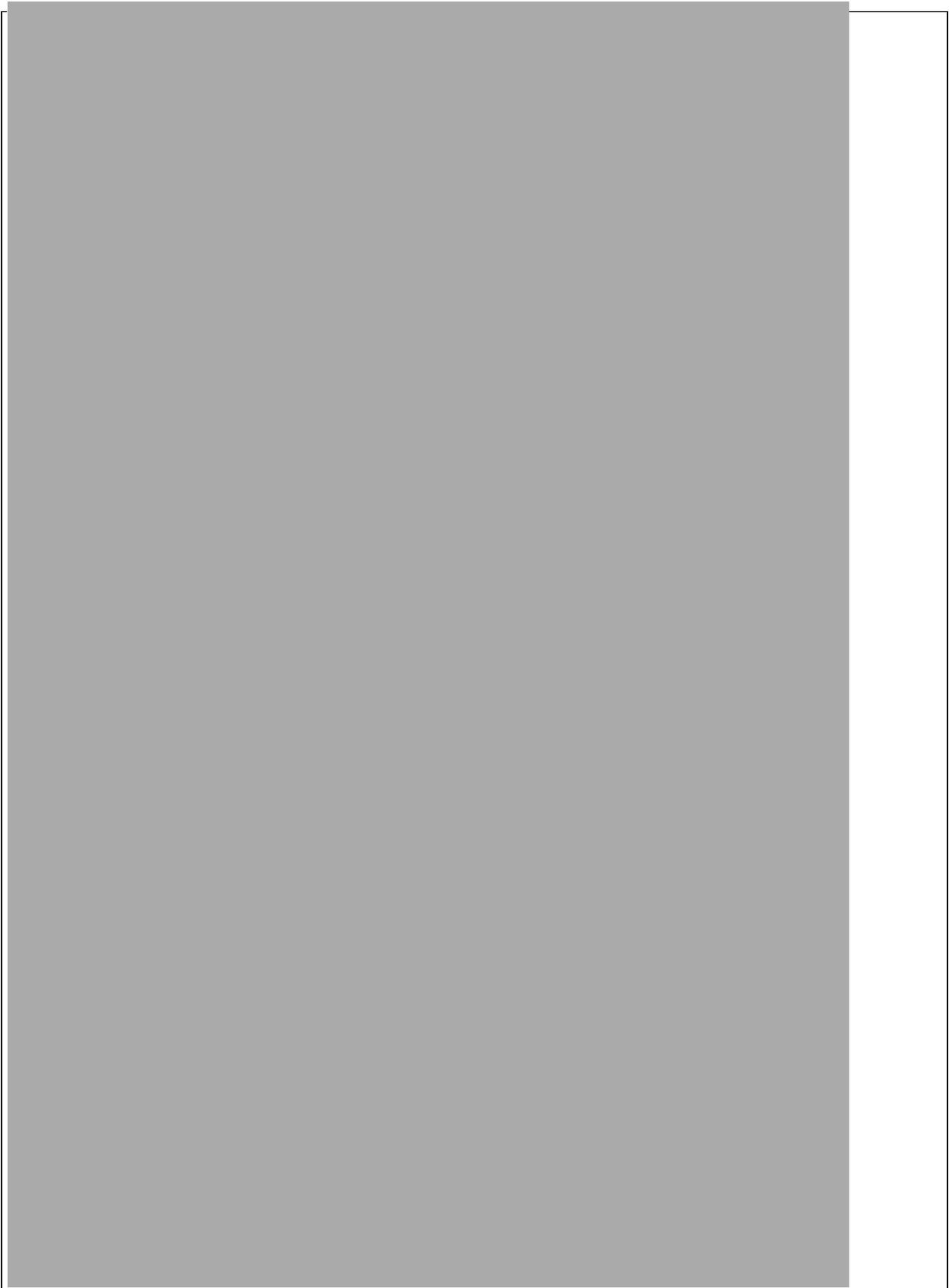
Score 1:

[Redacted content]



Score 2: Satisfactory



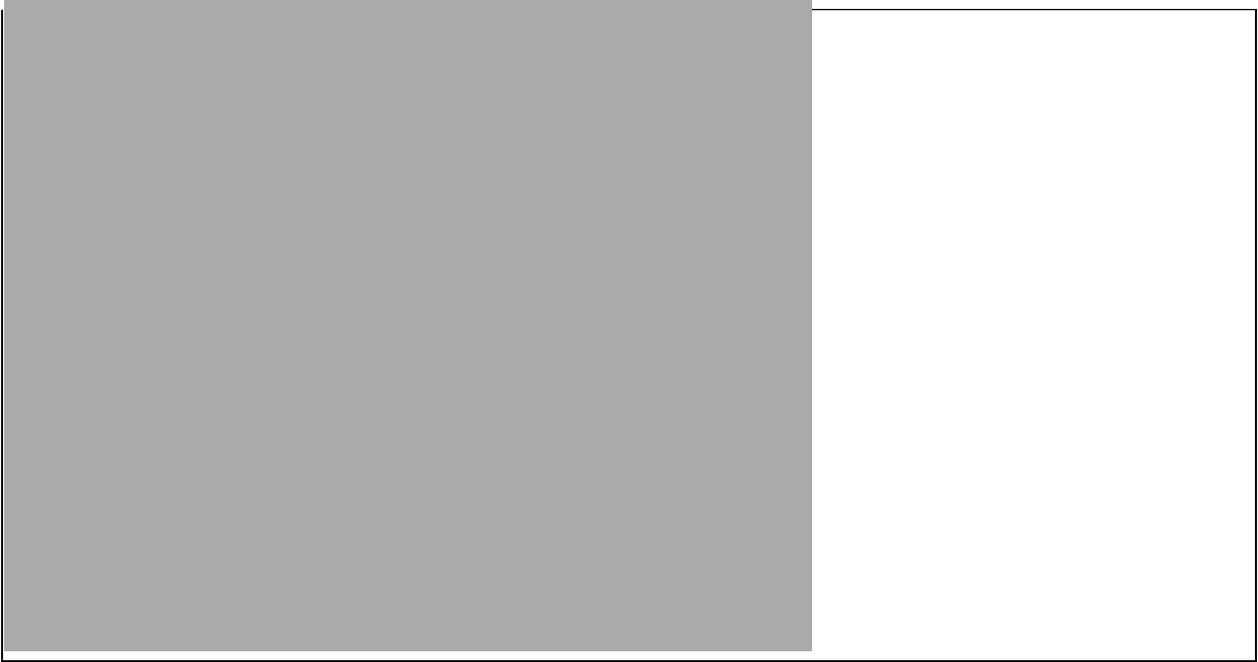


Score 3: Partial



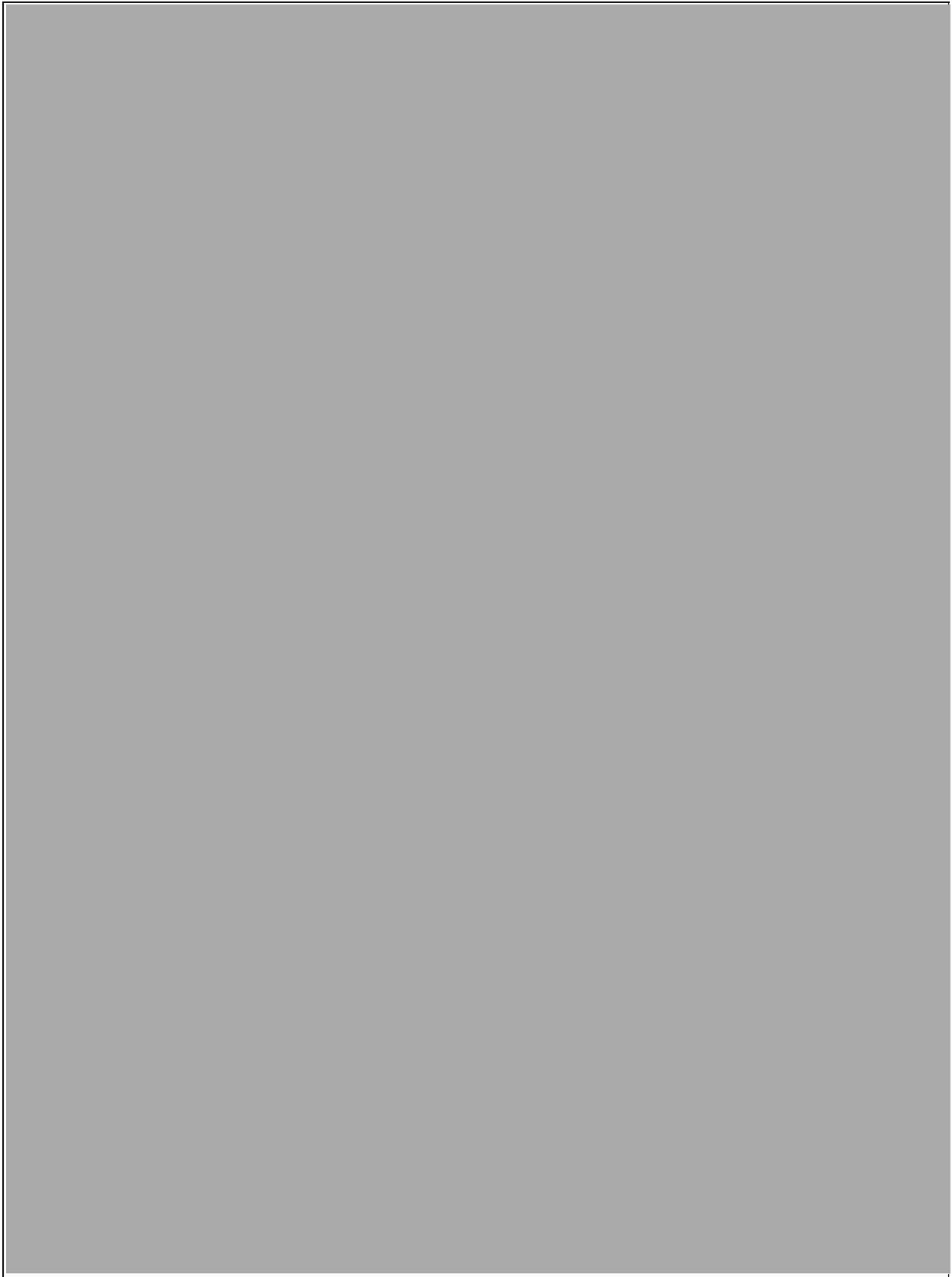
Score 4: Minimal





Score 5: Incorrect/ Off task





1b



2b



11 b, c, d



15)



18)



19)

