Psychometric Report for the Early Fractions Test (Version 2.2) Administered With Third- and Fourth-Grade Students in Spring 2017

Robert C. Schoen
Xiaotong Yang
Sicong Liu
Insu Paek

SECURE VERSION

Suggested citation: Schoen, R. C., Yang, X., Liu, S., & Paek, I. (2017). *Psychometric report for the Early Fractions Test (version 2.2) administered with third- and fourth-grade students in spring 2017*. (Research Report No. 2017-11). Tallahassee, FL: Learning Systems Institute, Florida State University.

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).

Psychometric Report for the Early Fractions Test (Version 2.2) Administered with Third- and Fourth-grade Students in Spring 2017

Robert C. Schoen

Xiaotong Yang

Sicong Liu

Insu Paek

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. Xiaotong Yang and Sicong Liu collaborated on 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. Charity Bauduin provided valuable assistance with the style and format of the final report.

Catherine Lewis, Kevin Lai, Kristy Farina, 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. Without them, this work is not possible.

Table of Contents

List of Appendices

List of Tables

List of Figures

List of Equations

Executive Summary

The Early Fractions Test v2.2 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 1,229 third- and fourth-grade students in spring 2017 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 v2.2 is to serve as a post-intervention measure of student learning outcomes 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, item difficult estimates, test information, and the distribution of student ability estimates. The results of these analyses are largely consistent with the findings from the analysis of data from the previous version of the Early Fractions Test (Schoen, Liu, Yang, & Paek, 2017).

Description of the Test

The Early Fractions Test v2.2 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). It assesses elementary students' understandings of fundamental conceptions of fractions, including *partitioning and iterating*, *referent unit*, *magnitude comparison*, *fractions as number on a number line*, and *operations on fractions*. The test form contains 20 numbered items prompting up to 27 individual responses from the test taker, with seven of them using a selected-response format and 20 using a constructed-response format.

Sample and Setting

The Early Fractions Test v2.2 was administered with a sample of 1,229 third- and fourth-grade students in six U.S. states in spring 2017. 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. The tests were administered by the students' classroom teachers and scored by research project staff at Florida State University.

Results

Item Diagnostics and Scoring

Item diagnostics and calibration accounting resulted in the collapsing of the 27 individual responses (or non-responses) to a total of 18 independent items. All the 27 responses contributed to the final 18-item scale.

Initial screening of the items used an approach based on classical test theory (CTT). Item difficulty indices for the 18 items in the final scale ranged from .21 to .94. The lowest item-rest correlation coefficient was .28. All the other items had item-rest correlation coefficients between .37 and .68, suggesting that the items used in the final scale generally had good discriminative power.

Dimensionality

To investigate the dimensionality of the test data, we performed Exploratory Factor Analysis and Parallel Analysis using the final-scale (18-item) format. Results of these analyses suggested a single dominant factor in the Early Fractions Test v2.2 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 2-parameter logistic model, 3-parameter logistic model (to adjust for student guessing), and generalized partial credit model 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 used *expected a posteriori* (EAP) estimator.

Findings from IRT analyses indicated that the item discrimination indices ranged from 0.56 to 2.52 (*M* = 1.57, *SD* = 0.56). The item difficulty indices ranged from −2.19 to 1.22 (*M* = −0.34, *SD* = 0.99). The discrimination index was greater than 0.50 for each of the 18 items, and 15 of the items had discrimination indices above 1.00. Eleven items had item difficulty values below 0.00, and seven items had item difficulty values above 0.00.

Students' EAP theta estimates ranged from −2.38 to 1.96. The skewness statistic was −0.09, and the kurtosis statistic was −0.41.

Reliability and Test Information

Using a CTT approach, coefficient α and standard error of measurement (SEM) were calculated to be .84 and 2.40, respectively. Additionally, test information and conditional standard error of measurement (CSEM) were generated through an IRT-based approach. The highest test information and the lowest CSEM occurred when the person ability (i.e., θ) was approximately −0.40. The person ability estimate was associated with larger test information and lower CSEM for the person ability estimates between −1.60 and 0.80 on the θ scale and was associated with smaller test information and higher CSEM (i.e., higher CSEM) for the person ability estimates greater 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 different from a normal distribution. Using the EAP method, the theta estimates for the students in the sample ranged from −2.38 to 1.96 (*M* = 0.00, *SD* = 0.93). The skewness and the kurtosis statistics for the sample distribution were −0.09 and −0.41, respectively.

Discussion and Conclusions

In summary, we found that the Early Fractions Test v2.2 measures a dominant factor, supporting unidimensionality in the data. Reliability, test information, and item discrimination estimates appear to fit the intended purpose of the test. Evaluation of the structural validity of the resulting 18-item scale supports the assertion that the Early Fractions Test v2.2 meets or exceeds common standards for educational and psychological measurement for its stated purpose.

1. Introduction

Early Fractions Test v2.2 is designed to assess elementary students' understandings of fundamental conceptions of fractions, including *partitioning and iterating*, *referent unit*, *magnitude comparison*, *fractions as number on a number line*, and *operations on fractions*. Whereas none of the test items involve decimal numbers (e.g., 3.20, 0.75), test items do involve the use of conventional fraction terminologies and/or notations (e.g., one-half, one-sixth, ½). The items on the test emphasize linear representations of fractions as well as symbolic notation (e.g., ½). The contents of the test correspond to that of the Common Core State Standards for Mathematics (CCSS-M; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) for third- and fourth-grade students. Linear representations of fractions are emphasized in the Early Fractions Test, but items also involve students reading and writing representations of fractions involving numeral-based, symbolic notation (e.g., 1/3, ¾) and performing operations on fractions presented in symbolic notation.

The Early Fractions Test v2.2 was administered by classroom teachers in spring 2017. The field-test data analysis reported in the current report is based on a sample of 1,229 third- and fourth-grade students in 63 schools in six states. Students of both grades completed an identical fractions test in the test-form format (see Appendix A).

Early Fractions Test v2.2 was used as an outcome measure of students' fractions learning in a multisite cluster randomized controlled trial investigating the individual and combined effects of teacher involvement with lesson study and a fractions resource kit on student learning. The purpose of the test is to serve as a measure of student learning outcomes in fractions so that the effect of different intervention conditions can be compared with one another to determine the effect of the various components of the interventions on student outcomes. The current report is centered on scoring and data modeling of the data generated in the Early Fractions Test v2.2.

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; IES/NCES, 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).

Table 1.1 shows the allocation of test items according to content standards in the test blueprint. Because the original test items (i.e., test form) was reconfigured based on analyses described in later sections of this report, the number of items in the final test (i.e., final scale) differed from the number of original test form items. Explanations of the discrepancies are provided in chapters two and three of this report.

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	6	4
Magnitude Comparison	2	2
Operations on Fractions and Problem Solving	5	3
Partitioning and Iterating	11	6
Referent Unit	3	3
Total Number of Items	*27*	*18*

Note. Test Form = the test items in the original fraction test; Final Scale = the adjusted index numbers (with the symbol * to help differentiate from test-form item numbers) of all the individual responses in the statistical analyses.

2. Initial Item Review

Early Fractions Test v2.2 consists of 20 items in its original format when presented to student participants. The 20 items generated a total of 27 fraction-related responses. Six of these 27 items are presented to the test takers in a selected-response (i.e., multiple-choice) format, while the remaining 21 items are presented in a constructed-response format. The discrepancy between 20 and 27 is due to the fact that several items (i.e., items 1, 2, 11, 12 in the test form) were testlets that required more than one response.
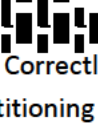
Although the contents of this report are presented in a linear-sequential manner, the actual psychometric operations were achieved through an interactive, iterative, and overlapping process. For instance, recoding the 20 test-form items into the 18 final-scale items was informed by the polychoric correlations between test form items 7–9.

At the stage of data entry, all the 27 responses were coded as dichotomous variables. Then, dichotomous variables indexed under the same test item were added together to generate a polytomous variable, resulting in 20 item-level variables. For instance, item 1 in the original test requires two responses that were coded dichotomously during data entry. Later, these two dichotomous variables were added to form a polytomous variable ranging from 0 to 2. Recoding dichotomous variables into polytomous variables helps address concerns about local dependence of items when applying item response theory (IRT) models to scoring students' latent abilities (de Ayala, 2009).

Another adjustment of item variables was performed later based on statistical reasons explained in section 3.3 of this report. That is, the 20 item variables were again recoded into 18 item variables by combining items 7, 8, and 9. To clarify the scoring of the items, the 27-, 20-, and 18-variable coding format was labeled data-entry, test-form, and final-scale, respectively.

Table 2.1 shows details regarding the test blueprint, as well as correspondence of items in the data-entry, test-form, and final-scale format. The final-scale items could also be distinguished from test-form items by having an * after their item numbers. For example, item 1* represents the first item in the final-scale format, whereas item 1 stands for the first item in the test-form format.

Table 2.1. Detailed Test Blueprint for the Spring 2017 Early Fractions Test v2.2, Split by Test Formats

Item description	Data entry #	Test form #	Final scale #
Fractions as Number on a Number Line			
Rabbit Problem Part 1	1a	1	1*
Rabbit Problem Part 2	1b	1	1*
Polar Bear Problem Part 1	2a	2	2*
Polar Bear Problem Part 2	2b	2	2*
Determine the point on the NL ▮ MC	15	15	13*
Determine the point on the NL ▮ MC	16	16	14*
Magnitude Comparison			
▮ gallon vs. ▮ gallon (Pretest CR/Posttest MC)	5	5	5*
Determining the greatest fraction (MC)	6	6	6*
Operations on Fractions and Problem Solving			
2 people sharing ▮ yard	4	4	4*
▮▮▮▮	7	7	7*
▮▮▮	8	8	7*
▮▮▮▮	9	9	7*
Correctly Partitioned ▮▮ on NL	10	10	8*
Partitioning and Iterating			
Part of a Referent Unit ▮)	3	3	3*
▮ of the shaded ribbon	11a	11	9*
Shade ▮	11b	11	9*
Shade ▮	11c	11	9*
Shade ▮	11d	11	9*
Iterating unit fraction box (▮ pieces of ▮ is $\frac{\square}{\square}$)	12a	12	10*
Iterating unit fraction box (□ pieces of ▮ is $\frac{▮}{0}$)	12b	12	10*
Iterating unit fraction box ($\frac{\square}{▮}$ = ▮)	12c	12	10*
Fourths in ▮▮▮	13	13	11*
Fourths in ▮	14	14	12*
Joe's walk	20	20	18*
Referent Unit			
Jose and Ella's pizzas	17	17	15*
Determining Referent Unit from ▮	18	18	16*
Draw ▮	19	19	17*
Total Number of Items	27	20	18

Note. Question Description = description of the fraction questions; Data Entry # = the index numbers of data entry (dichotomous) variables that correspond to all the 27 responses tapped by the test; Test Form # = the index numbers of all the items in the original fraction test (see Appendix A); Final Scale # = the adjusted index numbers (with the symbol * behind to help differentiate from test-form item numbers) of all the items in the statistical analyses.

3. Data Entry and Item-level Scoring

3.1. Sample

The Early Fractions Test v2.2 was administered with 1,229 third- and fourth-grade students representing six U.S. states in spring 2017. The students were recruited through their teachers who volunteered to join a randomized-controlled trial. The trial was designed to investigate the effects of fractions resource toolkits and lesson study on student learning.

The students took the test in a paper-pencil format. All students completed the same version of the test (Appendix A). The tests were administered by the students' teachers. Project staff at the Florida State University mailed the test copies, together with manuals of administration instructions, to participating schools, whose employees (e.g., students' classroom teachers) administered the test with students with positive assent and parental consent according to the administration manual. The test administration manual is provided in Appendix B. Test administrations took place during a period spanning March 3, 2017 through June 15, 2017.

Among the 1,229 students, five students had missing responses. The five cases were deleted in all the analyses of this report based on the following reasons. First, the proportion of the missing cases is small (i.e., 0.41%). Second, because the report generated student ability estimates based on *item-response theory*, and students' person ability estimation of complete cases cannot be completely comparable to that of the five missing cases. Lastly, the inclusion of any student who has missing data would result in varying sample sizes across different analyses (which tend to have different treatment of missing values). Deleting the five missing cases in this report helps maintain accuracy and consistency of the reported information with a low cost of information loss. Therefore, the final sample size is 1,224. Table 3.1 shows the demographic information of the final sample adopted in this report.

Table 3.1. Demographic Characteristics of the Students (n = 1,224) in the Spring 2017 Field-test of the Early Fractions Test v2.2

Characteristic	Number (Proportion of sample)
Language	
ELL	158 (.13)
Non-ELL	782 (.64)
Unknown	284 (.23)
Grade level	
Third	499 (.41)
Fourth	555 (.45)
Unknown	170 (.14)
Gender	
Male	459 (.38)
Female	513 (.42)
Unknown	252 (.21)
State	
FL	661 (.54)
CA	142 (.12)
IL	162 (.13)
NY	61 (.05)
CO	17 (.01)
IN	11 (.01)
Unknown	170 (.14)

Note. ELL= English-Language Learner. Gender and ELL 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.

3.2. Data Entry and Verification Procedures

A team of three 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. 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.

At the point of data entry, codes for correct (1), incorrect (0), missing (8), and did not solve (9) were used. Item-level data coded as *missing* (8) were known to be never presented to the student. This was known to have happened in a few occasions where students were given the wrong form, or where a form was missing a page. The *did not solve* (9) code was used when the complete form was given to the student, but the student did not indicate any response to the item. As a result, we recoded the 9s in the initial data set to be incorrect responses, and we considered 8s in the data set to be system missing.

To verify that data entry and scoring guidelines were being conducted consistently across data entry personnel, a random sample of seven schools (representing 10% 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. Comparison of the initial data entry produced an agreement of 98% for scored item level data. The most frequent discrepancies were found on items 18 and 19 on the original test form. To alleviate this discrepancy, research assistants met to score these items as a group. At least two raters viewed each item, and any disagreements were discussed and recorded as notes in the scoring criteria. Once these corrections were made, the scored data agreed at a rate greater than 99% between the two records when compared on each item.

3.3. Item Scoring

The test developers provided an answer key and scoring rubric for the test, which were used to determine the correctness of item responses. The scoring rubric is provided in Appendix C.

As previously explained in Section 2, 27 dichotomous data-entry variables were first recoded into 20 test-form variables (that exactly match the item indexing in the original test). This recoding is necessary, because several test-form items require more than one response and, if not recoded, these items post threat to the local independence assumption for IRT-based modeling. To score the test-form items requiring more than one response, we generated polytomous variables by summing relevant dichotomous variables under the same testlets.

After scoring each of the 20 test-form items, we further adjusted the item coding in a special case due to statistical reasons. The case is related to test-form items 7, 8 and 9, which are three fill-in-the-blank (constructed-response) items. Although each of the three items was dichotomously scored at the beginning, they were combined into one final-scale item represented by a polytomous variable (i.e., item 7*) based on the following reasons. First, the three items were arranged sequentially in the test, and they were introduced by a shared direction (see Appendix A). Second, high polychoric correlations between any two of these three items were evident (i.e., .98 for items 7 & 9, .96 for items 7 & 8, and .94 for items 8 & 9), which leads to a concern about item-dependency among the three items. Table 3.2 shows the details of the recoding process.

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	8*	0, 1
11	0, 1, 2, 3, 4	9*	0, 1, 2, 3, 4
12	0, 1, 2, 3	10*	0, 1, 2, 3
13	0, 1	11*	0, 1
14	0, 1	12*	0, 1
15	0, 1	13*	0, 1
16	0, 1	14*	0, 1
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-Format 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 the symbol * after the final-scale item number).

4. Dimensionality Analysis

4.1. Exploratory Factor Analysis

Exploratory factor analysis (EFA) was conducted to examine the dimensionality of the test using Mplus 8.0 (Muthén & Muthén, 1998–2017). Given the ordinal nature of the item response, and lack of symmetry in item responses, we conducted the analysis using weighted least square estimation method with mean and variance adjusted (WLSMV; Finney & Distefano, 2013), and the Geomin rotation method. The eigenvalues estimated by Mplus and the corresponding percentages of variation explained are displayed in Table 4.1. The first factor explained 46.67% of the variation. Figure 4.1 shows the scree plot for the eigenvalues. Based on the evidence, 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	8.40	46.67
2	1.17	6.50
3	0.99	5.50
4	0.86	4.78
5	0.82	4.56
6	0.71	3.94
7	0.71	3.94
8	0.61	3.39
9	0.58	2.89
10	0.56	3.22
11	0.50	2.78
12	0.43	2.39
13	0.39	2.17
14	0.34	1.89
15	0.31	1.72
16	0.24	1.33
17	0.23	1.28
18	0.18	1.00

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.
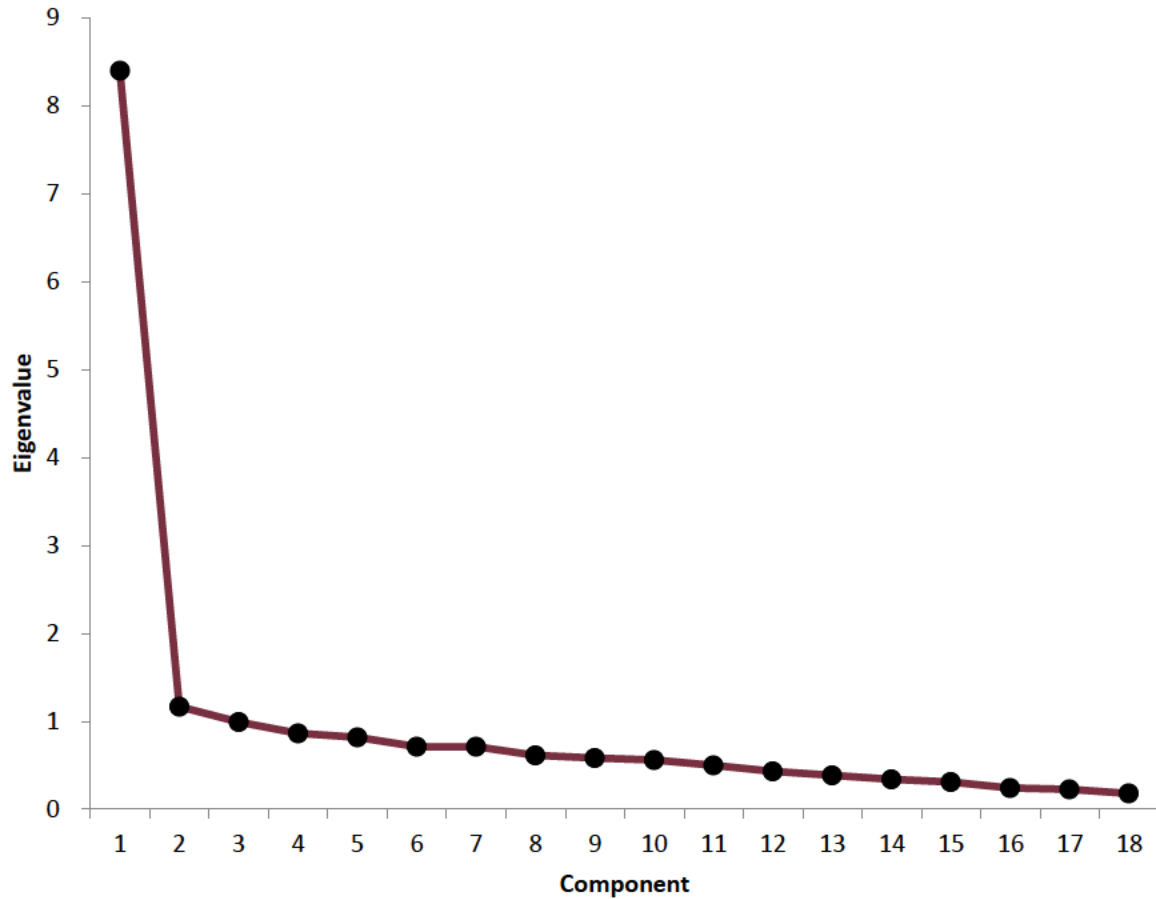
Figure 4.1. Scree plot for the eigenvalues estimated from Mplus.

4.2. Parallel Analysis

We also performed parallel analysis (PA) to check the dimensionality of the test. The *psych* (Revelle, 2017) package in R 3.4.0 (R Core Team, 2017) was used to perform PA. Parallel analysis (PA) is a procedure to help decide how many components to retain 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 results of the PA supported unidimensionality. That is, the test measures a single, dominant construct.

5. Classical Testing Theory (CTT) Analyses

We first analyzed the data based on classical testing theory (CTT). All the results subsequently presented were obtained with SPSS 22.0 (IBM corp., 2013). The results included (a) the distribution of the observed test scores, (b) item difficulty and discrimination, and (c) reliability and standard error of measurement.

5.1. Distribution of the Observed Test Score

Figure 5.1 displays the distribution of the observed total test score in the sample. Note that although the final-scale format had 18 items, the observed test scores ranged from 0 to 27, because there were some polytomous items (i.e., items 1*, 2*, 7*, 9*, 10*). The mean of the total test score was 17.74, and the standard deviation was 5.99. The median of the total test score was 19.00. The skewness statistic was –0.57, and the kurtosis statistic was –0.43.
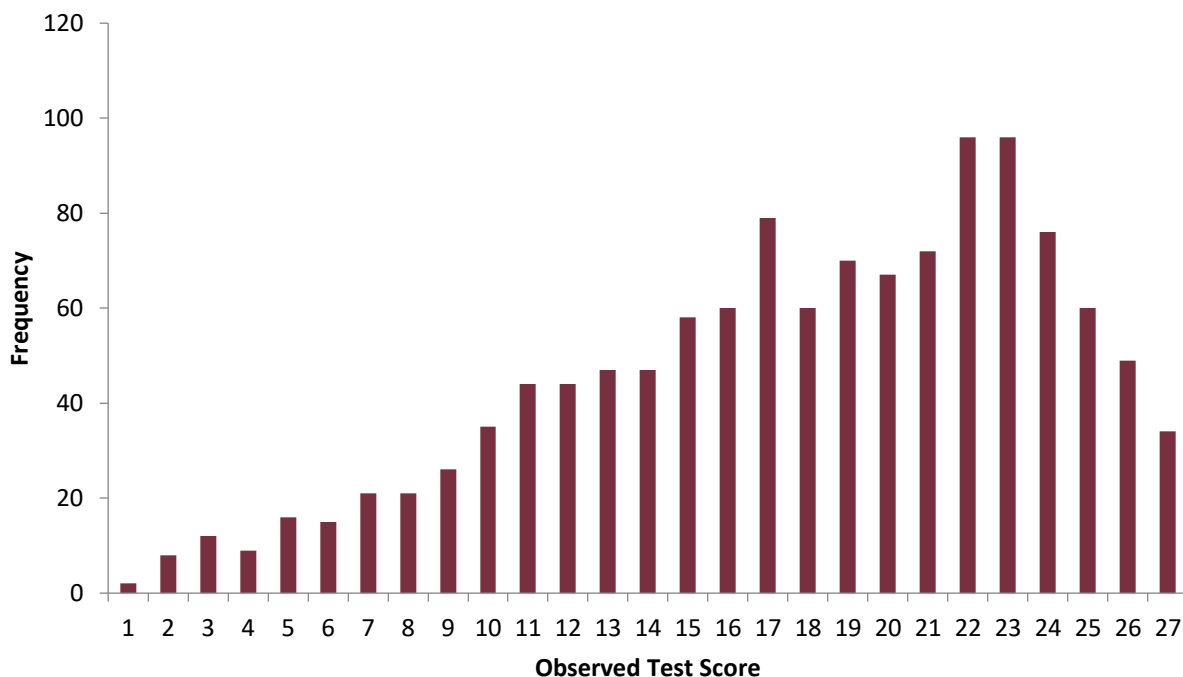
Figure 5.1. Bar graph depicting the distribution of the observed test score in the final-scale format.

5.2. Item Difficulty & Discrimination

For both dichotomous and polytomous items, the item difficulty indices could be calculated based on Equation 1 (McDonald, 1999). Note that when the items were dichotomously coded, the values of *p* are equivalent to the proportion of correct answers for that item.

$$p = \frac{\text{Item}_{\text{Mean}} - \text{Item}_{\text{Min}}}{\text{Theoretical Score Range}} \qquad (1)$$

where p is the symbol of the item difficulty index.

Table 5.1 shows the mean score, standard deviation, difficulty and discrimination indices for each of the final-scale items. The item difficulty indices varied from .21 (item 4*) to .94 (item 3*). The item discrimination indices (i.e. item-rest correlation coefficients) varied from .28 (item 3*) to a maximum of .68 (item 10*).

Table 5.1. Item Difficulty and Discrimination from CTT Analyses

Final-scale item #	M	SD	p	Item-rest r
1*	1.54	0.76	.77	.46
2*	1.33	0.84	.67	.58
3*	0.94	0.24	.94	.28
4*	0.21	0.41	.21	.38
5*	0.83	0.38	.83	.40
6*	0.77	0.42	.77	.47
7*	2.23	1.22	.74	.47
8*	0.46	0.50	.46	.39
9*	3.44	0.99	.86	.45
10*	2.11	1.07	.70	.68
11*	0.85	0.36	.85	.52
12*	0.33	0.47	.33	.51
13*	0.72	0.45	.72	.51
14*	0.53	0.50	.53	.54
15*	0.30	0.46	.30	.37
16*	0.29	0.46	.29	.38
17*	0.34	0.47	.34	.47
18*	0.54	0.50	.54	.52

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 the symbol * 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.

Tables 5.2 and 5.3 show the distribution of item difficulty and item discrimination for the 18 items used in the final scale. A wide range of item difficulty appears to be represented on the test (for this particular sample). The item discrimination estimates are greater than .40 for the majority of the items, but the item-rest correlation coefficients fall between .20 and .40 for five of the items, and none of the items have item-rest correlation coefficients greater than .70.

Table 5.2. Distribution of CTT-based Item Difficulty (p-values) Estimates for Items Used in the Final Scale

p-value	Number of items
>.90	1
.80 – .89	3
.70 – .79	5
.60 – .69	1
.50 – .59	2
.40 – .49	1
.30 – .39	3
.20 – .29	2
.10 – .19	0
<.09	0
Mean	0.60
Median	0.69
Standard Deviation	0.23

Table 5.3. Distribution of CTT-based Item Discrimination (Item-Rest r) Point Estimates for Items Used in the Final Scale

Item-Rest r	Number of items
.90–1.00	0
.80–.89	0
.70–.79	0
.60–.69	1
.50–.59	6
.40–.49	6
.30–.39	4
.20–.29	1
.00–.10	0
Mean	0.47
Median	0.47
Standard Deviation	0.09

Note. Mean and Median look identical due to rounding errors.

5.3. Coefficient Alpha & Standard Error of Measurement

We calculated Coefficient α (Cronbach, 1951) as one way to estimate the test reliability. The Coefficient α of the test was .84. We also calculated the standard error of measurement (SEM) of the test. The scale had a variance of 35.86. SEM was calculated to be 2.40 based on Equation 2, where σ^2 is the test variance, and ρ_{XX} is the Coefficient α of the test.

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

6. Item Response Theory (IRT) Analyses

6.1. Model Description

We conducted the IRT analyses using flexMIRT 3.5 (Cai, 2017). For the constructed-response items that were scored dichotomously (i.e. items 4*, 11*, 12*, 15*, 16*, 17*, and 18*), the two-parameter (2PL) model was used. For the 2-option, 4-option, and 5-option multiple-choice items that were scored dichotomously (i.e. items 3*, 5*, 6*, 8*, 13* and 14*), the three-parameter (3PL) model was used, which adjusted for guessing. For the polytomously scored items (i.e. items 1*, 2*, 7*, 9* and 10*), the Generalized Partial Credit Model (GPCM) was used.

The formulas of 2PL model, 3PL model, and GPCM are shown below (de Ayala, 2009). Successful convergence was reached in the computation for the IRT analyses, and −2loglikelihood was 25483.02.

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

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

(3)

where

a_j is the discrimination index of item j (j = 1, 2,…,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 3PL model is presented in Equation 4,

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

(4)

where

a_j is the discrimination index of item j (j = 1, 2,…,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 GPCM is presented in Equation 5,

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

where

a_j is the discrimination index of item j (j = 1, 2,…,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 hth threshold, k is item category, $k \in \{0, 1, 2, …, m_j\}$.

6.2. Item Difficulty and Discrimination

Table 6.1 presents descriptive statistics of item difficulty and item discrimination indices of the 18 items. The item discrimination indices ranged from 0.56 to 2.52 (M = 1.57, SD = 0.56). The item difficulty indices ranged from −2.19 to 1.22 (M = −0.34, SD = 0.99). Tables 6.2, 6.3, and 6.4 present parameter estimates for the items using 2PL, 3PL, or GPCM models, respectively. Figure 6.1 displays the item discrimination estimates of each item. The discrimination indices for all the 18 items were greater than 0.50, and 15 of the items had discrimination indices above 1.00. Figure 6.2 displays the item difficulty estimates for all the items. Eleven items had b values below 0.00, and 7 items had b values above 0.00.

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

	M	*SD*	Min	Max	Skewness	Kurtosis
a	1.57	0.56	0.56	2.52	−0.07	−0.54
b	−0.34	0.99	−2.19	1.22	−0.11	−0.85

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)
4*	1.55(0.13)	1.22(0.08)
11*	2.48(0.22)	−1.27(0.07)
12*	2.04(0.16)	0.57(0.05)
15*	1.15(0.10)	0.95(0.09)
16*	1.30(0.11)	0.89(0.08)
17*	1.60(0.13)	0.60(0.06)
18*	1.63(0.12)	−0.13(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 the symbol * 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*	1.37(0.19)	−2.19(0.27)	0.27(0.10)
5*	1.88(0.30)	−0.73(0.20)	0.41(0.08)
6*	2.03(0.26)	−0.71(0.13)	0.22(0.06)
8*	1.69(0.24)	0.52(0.10)	0.17(0.04)
13*	2.12(0.23)	−0.55(0.10)	0.16(0.05)
14*	2.52(0.28)	0.10(0.07)	0.12(0.03)

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 the symbol * 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*	0.84(0.08)	−1.20(0.09)	−0.81(0.16)	0.81(0.16)		
2*	1.13(0.09)	−0.63(0.06)	−0.25(0.09)	0.25(0.09)		
7*	0.56(0.05)	−1.00(0.07)	−4.27(0.53)	1.87(0.42)	2.40(0.30)	
9*	0.70(0.05)	−1.86(0.11)	0.36(0.28)	−0.62(0.28)	0.31(0.21)	−0.05(0.14)
10*	1.61(0.12)	−0.74(0.05)	0.59(0.06)	−0.30(0.07)	−0.29(0.06)	

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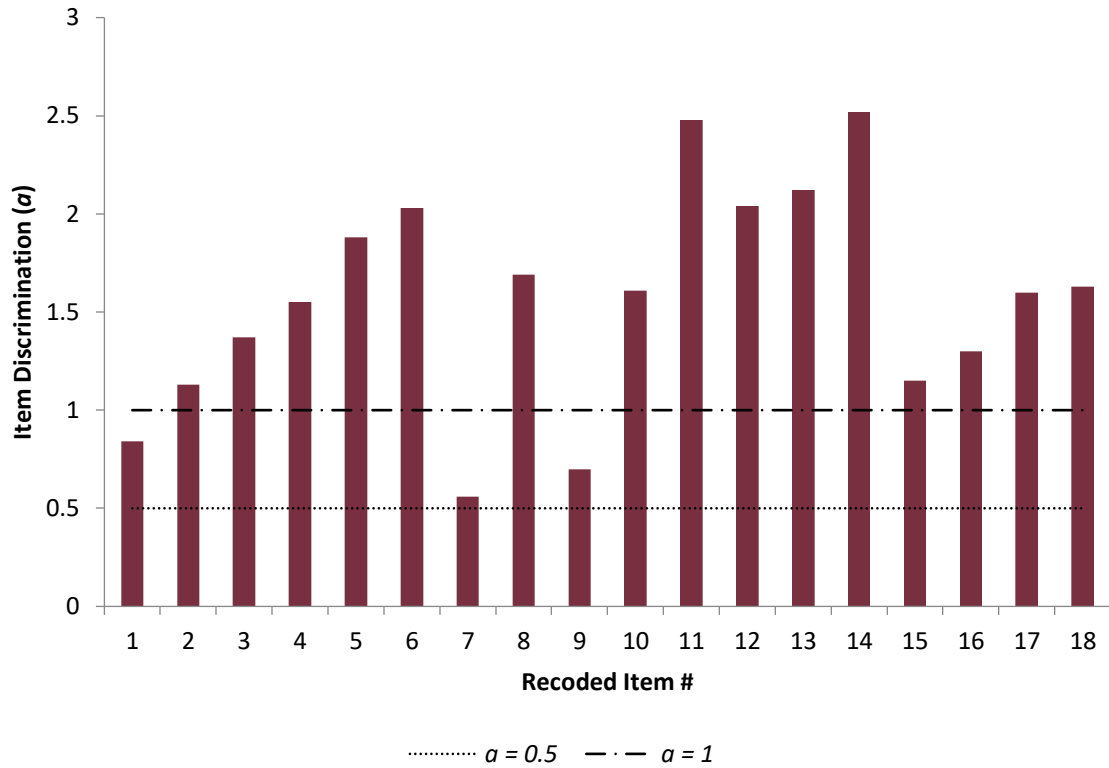
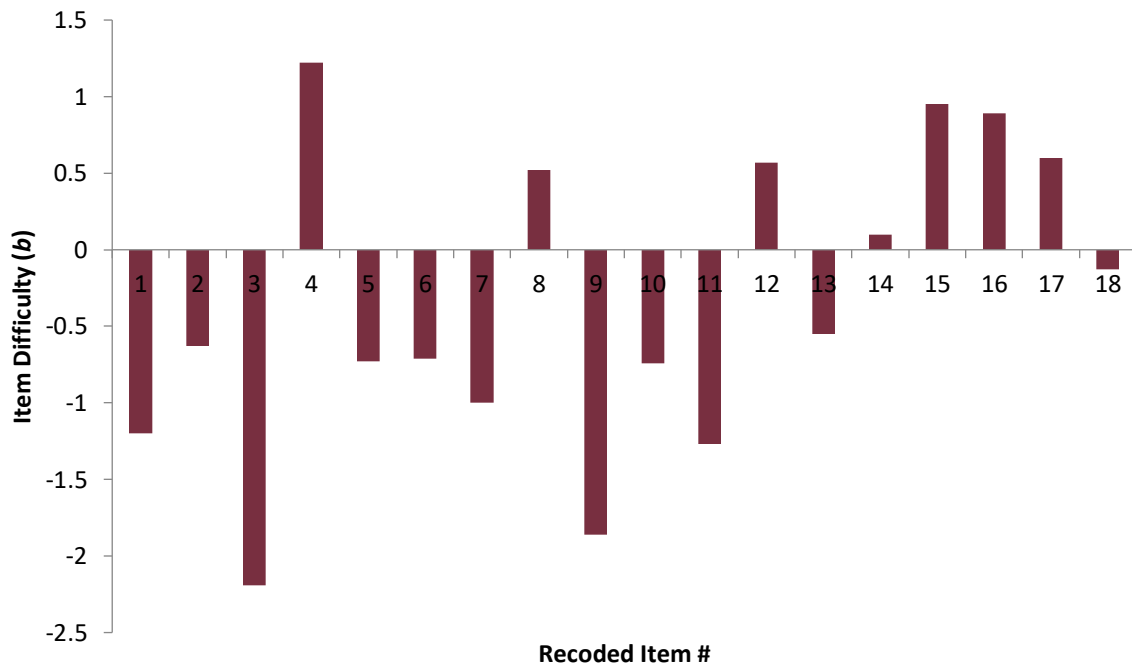Figure 6.1. Item discrimination estimate (a) of each final-scale item.

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 test information curve and the test conditional standard error of measurement (CSEM). Equation 6 shows the formula to calculate CSEM given person ability (de Ayala, 2009). In the equation, I is the test information function given person ability, and θ is the person ability.

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

Based on Figure 6.3, the person-ability (i.e., θ) estimates near −0.40 on the theta scale are associated with the largest test information and the smallest CSEM. In addition, Figure 6.3 suggests that the person-ability estimate was related to the smallest CSEM when it ranged between −1.60 and 0.80, and it was related to the largest CSEM when it was greater than 2.00.
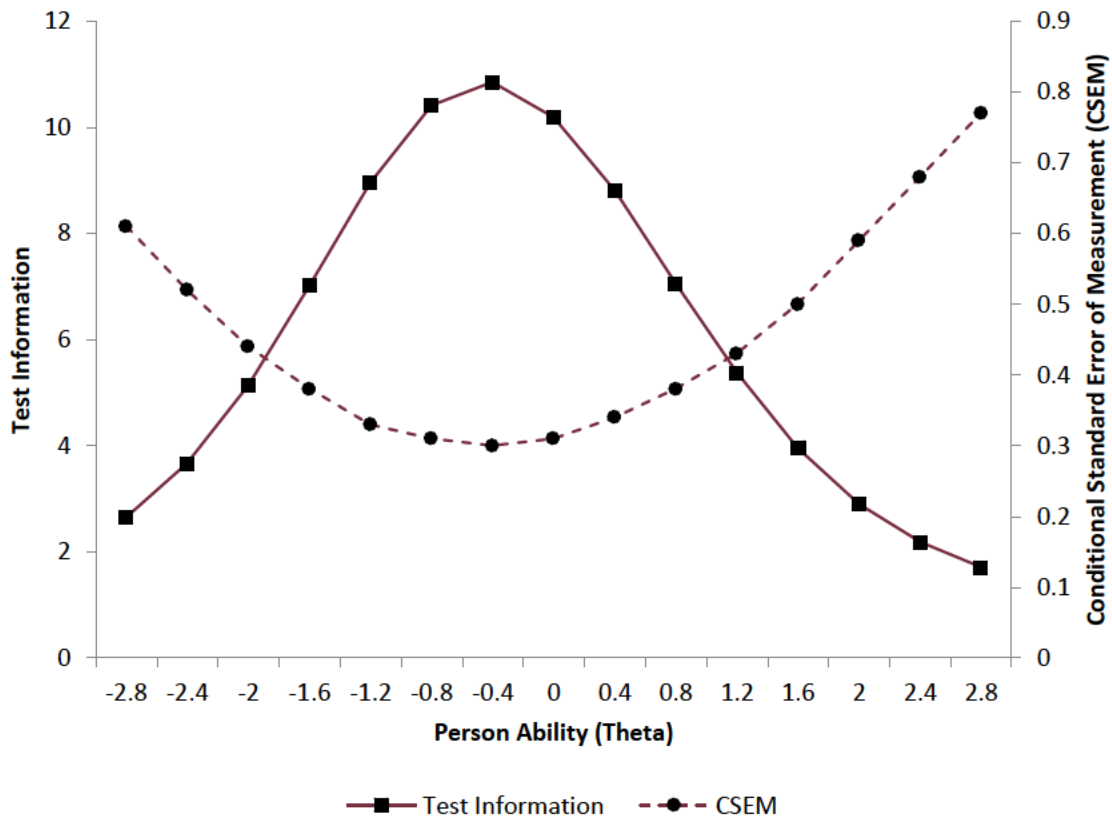

Figure 6.3. Test information curve and conditional standard error of measurement (CSEM) for the final-scale format.

Figure 6.4 shows the distribution of the theta estimation (*M* = 0.09, *SD* = 1.40) using MLE. The skewness and kurtosis statistics were 1.15 and 4.94, respectively. No students had zero scores. However, 34 students had perfect scores, including 12 third grade students, 20 fourth grade students, and 2 students whose grade-level information was missing. As shown in Figure 6.4, spikes at the higher end of the horizontal axis existed, because some students had perfect scores for the test. When students had perfect scores, their MLE estimates were not available.

We also used *expected a posteriori* (EAP) method for the theta estimation. Using EAP method, Figure 6.5 shows the distribution of the theta estimation, which ranges from −2.38 to 1.96 (*M* = 0.00, *SD* = 0.93). The skewness and the kurtosis statistics were −0.09 and −0.41, respectively.
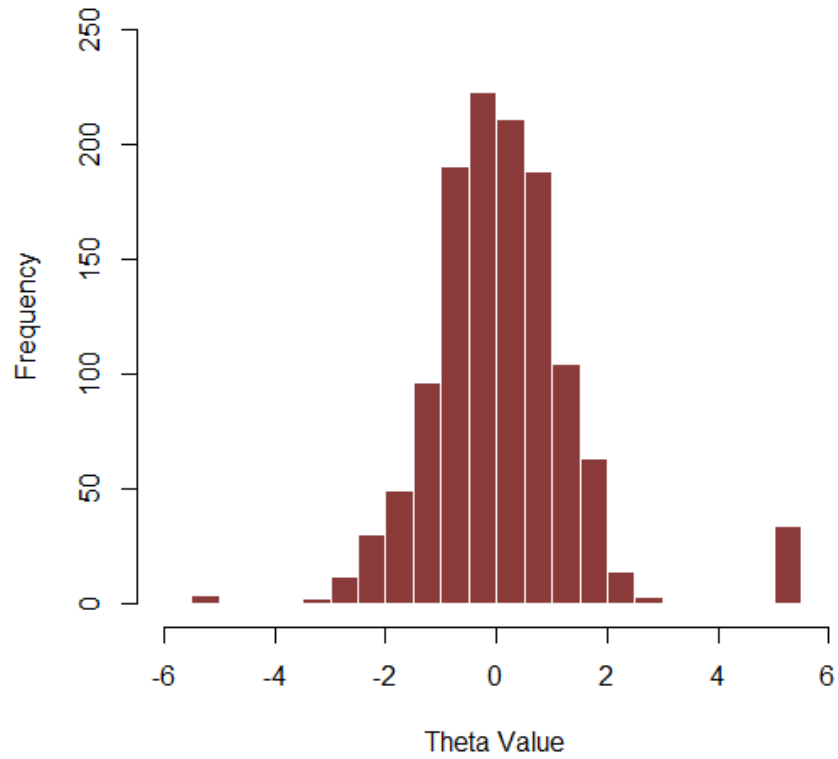
Figure 6.4. Person abilities (i.e., ϑ) estimated by maximum likelihood estimation (MLE).
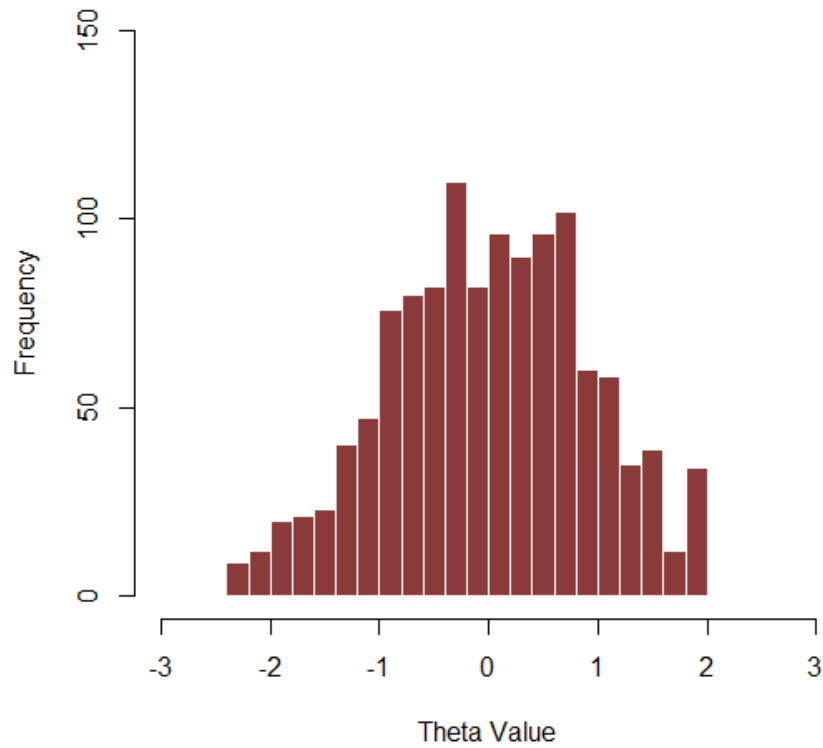

Figure 6.5. Person abilities (i.e., ϑ) estimated by expected a posteriori (EAP)

7. Discussion and Conclusions

The current report provides substantial evidence in support of substantive and structural validity (Flake, Pek, & Hehman, 2017). In summary, we found that the Early Fractions Test v2.2 appears to measure a single, dominant trait, supporting an assumption of unidimensionality in the data. Reliability and item discrimination estimates appear to be sufficiently high. Evaluation of the structural validity of the resulting 18-item scale supports the assertion that the Early Fractions Test v2.2 meets or exceeds common standards for educational and psychological measurement for its stated purpose.

During data analysis, we found that several items in the test potentially threaten the validity of the local-independence assumption. For instance, we found evidence of multicollinearity in three very similar items when they were modeled separately. Consequently, the three dichotomously scored original items were collapsed into a single, polytomous item with reasonable parameter estimates, and only the polytomous variable contributed to the final scale. Several other items were presented as testlets. Modeling the responses to these testlets as polytomous variables preserved the assumption of the local-independence and resulted in items with stable parameter estimates.

The overall difficulty level appears to fit the ability level of the third- and fourth-grade students in the sample, but the test may be improved if it were to further challenge those students with above-average ability levels by inclusion of more relatively difficult items. This is evidenced by the slightly skewed distribution of total scores in the sample, the high CSEM for students above .80 on the theta scale, and the absence of items with CTT-based difficulty estimates less than .20. Fourth-grade students performed better than third-grade students on the test overall, but approximately one-third of the students who received a perfect score were in the third grade.

We note a potential avenue for exploration in future work. In the present sample, the individual students were nested in classrooms in different schools. The analyses conducted for this report did not account for this multilevel structure. A more refined analysis could accommodate this structure.

The intended use of the Early Fractions Test v2.2 is to serve as a measure of student achievement in a randomized controlled trial. The trial was designed to examine the effect of four different interventions on student mathematics achievement. Lewis and Perry (2017) used the original version of the test and scored it using an approach based on classical test theory. Their results provide some evidence that the test may be sufficiently sensitive to detect a treatment effect. Results concerning the ability of the test to detect a potential treatment effect using the spring 2017 data are not available at the time of publication of the present report.

References

Beckmann, S. (2005). *Mathematics for elementary teachers*. Boston, MA: Pearson Education.

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 strucuture 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.

Institute of Education Sciences/National Center for Education Statistics (IES/NCES). (2007). NAEP Questions Tool. Retrieved from http://nces.ed.gov/nationsreportcard/itmrlsx/landing.aspx

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.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141–151.

Hackenberg, A., Norton, A., Wilkins, J., & Steffe, L. (2009, April). *Testing hypotheses about students' operational development of fractions.* Paper presented at the Research Presession 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.

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.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, L. K. & Muthén, B. O. (1998–2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

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 spring 2016 (Research Report No. 2017-23). Tallahassee, FL: Learning Systems Institute, Florida State University.

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.

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.

Appendix A.The Early Fractions Test (Version 2.2) Form

The form in this appendix is identical to the form used on the Early Fractions Test v2.2. As a result, no headers or footers are used in this section of the report.
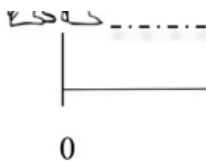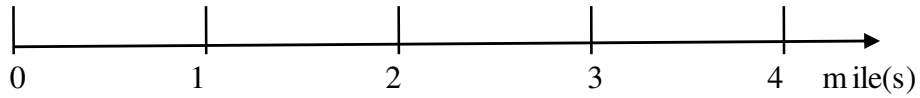
Student Fractions Questions (Post-Test)

2016-2017

Student Name: _____

Teacher Name: _____

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

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) ████████████████████████████████████
████████████████████████████████████
██████████

████████████████████████

```
├──────────┼──────────┼──────────┼──────────┼────────►
0          1          2          3          4    mile(s)
```

2) ████████████ $\frac{1}{4}$ ████████████████████████████
████████████████████████████████
██████

████████████████████████

```
├──────── . . . . . .
│
└──────────
0
```

3) What fraction of this shape is shaded?   Answer: _____

4) There are ████████████ $\frac{1}{4}$ ████████████████████████████████
████████████████████ how long is each string?

   Answer_____

5) Circle which is more: ████   █▌███

6) Which of the following fractions is the greatest?   Answer: _____

   A ▌   B ▌   C ▌   D ▌
      ▌   ▌   ▌   ▌
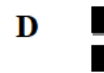
Write your answers to the following problems:

7) ■/■ + ■/■ = ?   Answer: _____

8) ■/■ + ■/■ = ?   Answer: _____

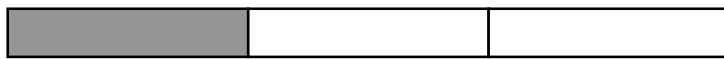9) ■/■ + ■/■ + ■/■ = ?   Answer: _____

10) Which number line is correctly divided into ■■?   Answer:_____
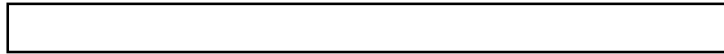
D) ├————
   0

4
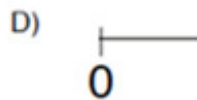
11) The whole bar shown below is 1 foot long.
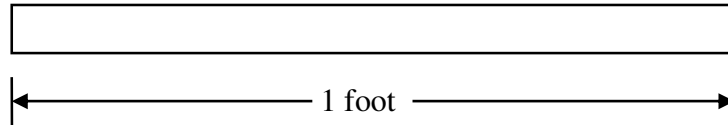
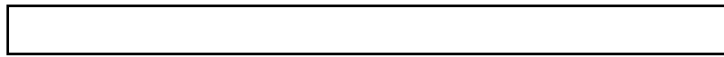How long is the shaded part of the bar shown below?

Answer: _____ foot

Shade in ▮ foot

Shade in ▮ foot
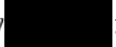

Shade in ▮ foot

12) Fill in the empty boxes with the missing numbers in the problems below:

Example: 1 piece of $\frac{1}{2}$ inch is $\dfrac{\boxed{1}}{\boxed{2}}$ inch

a) ▮▮▮▮ inch is $\dfrac{\square}{\square}$ inch

b) $\square$ pieces of ▮▮▮▮ inch

c) ▮▮▮▮   1

13) How many ▮▮▮▮ make a whole?   Answer: _____

14) How many ▮▮▮▮ cups?   Answer: _____

15) What number belongs in the box?   Answer_____

A ▮   B ▮   C ▮   D ▮   E ▮
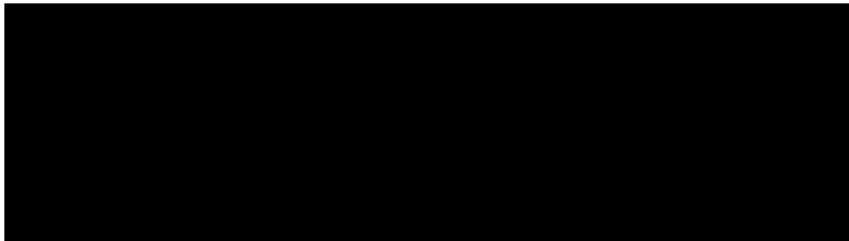
16) What number belongs in the box?   Answer_____

A

17) Think carefully about the following question. Write a complete answer. You may use drawings, words, and numbers to explain your answer. Be sure to show all of your work.
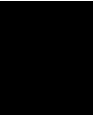
he same

18) The length of the bar shown below ▪ ▪ the whole bar. Draw how long the whole bar would be.

┌─────────────────┐
│   │
└─────────────────┘

19) The whole bar is shown below. Draw a bar that ▪ ▪ length of the whole bar.

┌───────────────────────────┐
│   │
└───────────────────────────┘

20) Joe walke ▪ le. How much farther must he walk to go 1 mile?

Appendix B. Administration Instructions

The form in this appendix is identical to the form used on the Early Fractions Test v2.2. As a result, no headers or footers are used in this section of the report.

Instructions for Administration of the *Student Fractions Questions* (Post-Test)

Overview

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

Materials Needed for Testing

The following materials are needed for the posttest:

- One copy of the *Student Fractions Questions* (Post-Test) for each student (provided)
- At least one sharpened pencil for each student

Testing

The *Student Fractions Questions* (Post-Test) is designed to be administered in a whole-class setting, with students completing the test independently. Students write their answers directly on the test. Give the post-test as you would other student tests. For example, have students space out desks or use privacy folders if that is what they usually do.

Please administer the post-test 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 post-test:

  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 post-test, 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 post-test will require approximately 30–40 minutes.

Submitting the *Student Fractions Questions* (Post-Test) Materials

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

Appendix C. Scoring Criteria

The forms in this appendix are identical to the forms used on the Early Fractions Test v2.2. As a result, no headers or footers are used in this section of the report.