2018

Psychometric report on the Knowledge for Teaching Elementary Fractions test administered to elementary educators in six states in fall 2016

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

Psychometric Report on the Knowledge for Teaching Elementary Fractions Test Administered to Elementary Educators in Six States in Fall 2016

Robert C. Schoen
Xiaotong Yang
Sicong Liu
Insu Paek

MAY 2018

Research Report No. 2018-12

SECURE VERSION

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., Yang, X., Liu, S., & Paek, I. (2018). *Psychometric report on the Knowledge for Teaching Elementary Fractions test administered to elementary educators in six states in fall 2016*. (Research Report No. 2018-12). Tallahassee, FL: Learning Systems Institute, Florida State University. https://doi.org/10.17125/fsu.1531453537

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 on the Knowledge for Teaching Elementary Fractions Test Administered to Elementary Educators in Six States in Fall 2016

Research Report No. 2018-12

Robert C. Schoen

Xiaotong Yang

Sicong Liu

Insu Paek

May 2018

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

Acknowledgments

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's coauthors.

Robert Schoen collaborated on the development of this form of the test, 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 item-response theory 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, Kevin Lai, Claire Riddell, and Robert Schoen developed the test, primarily through selection or adaptation of items drawn from other published sources. Claire Riddell created the Qualtrics-based version of the test with assistance from Amanda Tazaz. Claire Riddell managed the administration of the test. Robert Schoen and Claire Riddell managed the recruiting, enrollment, and consent processes. Kristy Farina managed the data entry and item-level scoring as a result of the adjudication process. She also assisted with preparation of the teacher demographic statistics and description of data entry and scoring criteria for the present report. Anne B. Thistle reviewed the copy for grammar, style, and formatting. Charity Bauduin provided valuable assistance with the style and format of the final report.

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.

Items on this test were borrowed or adapted from several different sources, including the Diagnostic Teacher Assessment in Math and Science project (DTAMS; Saderholm, Ronau, Brown, & Collins, 2010), the Learning Mathematics for Teaching project (LMT; Hill, Schilling, & Ball, 2004; LMT, 2004), Numeracy Development Projects (Ward & Thomas, 2015), and other publications (Beckmann, 2005; Newton, 2008; Norton & McCloskey, 2008; Schifter, 1998; Zhou, Peverly, & Xin, 2006). We are grateful to these people and organizations for sharing their intellectual property for research purposes.

We are especially grateful to the Institute of Education Sciences at the U.S. Department of Education for their support and to the elementary-school educators who agreed to participate in the study. Without them, this work is not possible.

Table of Contents

Acknowledgments 4
Executive Summary 11
Purpose Statement 11
Description of the Test 11
Sample and Setting 11
Results 12
Item Diagnostics and Scoring 12
Dimensionality 12
Item Response Theory Data Modeling 12
Reliability and Test Information 12
Distribution of Educator Ability Scores 12
Discussion and Conclusions 13
1. Introduction 14
1.1. Description of the Sample 14
1.2. Detailed Test Blueprint 15
2. Initial Item Review 17
3. Data and Scoring 18
3.1. Data Entry and Verification Procedures 18
3.2. Item Scoring 18
4. Dimensionality Analysis 20
5. Classical Testing Theory (CTT) Analyses 22
5.1. Distribution of the Observed Test Score 22
5.2. Item Difficulty & Discrimination 22
5.3. Coefficient α and Standard Error of Measurement 26
6. Item-Response Theory Analyses 27
6.1. Model Description 27
6.2. Item Difficulty and Discrimination 27
6.3. Test Information and Estimated Person Ability 28
7. Discussion 31

7.1. Substantive Validity 31
7.2. Structural Validity 31
7.2.1. Unidimensionality 31
7.2.2. Level of Difficulty for the Intended Population 31
7.2.3. Test Information 32
7.3. External Validity 32
7.4. Conclusions 32
References 33

List of Appendices

Appendix A. Sources of Assessment Items 35

Appendix B. Knowledge for Teaching Elementary Fractions Test 38

List of Tables

Table 1.1. Characteristics of Teachers in the Fall 2016 Field-Test Sample (N = 266) 15

Table 1.2. Test Blueprint for the Fall 2016 Knowledge for Teaching Elementary Fractions Test 16

Table 3.1. Missing Response Frequency in the Sample 19

Table 4.1. Eigenvalues Estimated from Mplus and Their Corresponding Percentages of Explained Variation 20

Table 5.1. Item Difficulty and Discrimination from CTT Analyses 23

Table 5.2. Distribution of CTT-based Item Difficulty (p-values) Estimates for Items Used in the Final Scale 24

Table 5.3. Distribution of CTT-based Item Discrimination (Item-Rest r) Point Estimates for Items Used in the Final Scale 24

Table 6.1. Parameter Estimates and Standard Errors for Final-Scale Items Modeled Using a Two-Parameter Model 28

Table 6.2. Parameter Estimates and Standard Errors for Final-Scale Items Modeled by Means of a Graded-Response Model 28

List of Figures

Figure 4.1. Scree plot of eigenvalues estimated from Mplus. 21

Figure 5.1. Distribution of the observed test scores in the final-scale format. 22

Figure 5.2. Item difficulty estimate (b) of each final-scale item. 25

Figure 5.3. Item discrimination estimate (a) of each final-scale item. 25

Figure 6.1. Test information curve and conditional standard error of measurement for the final-scale items. 29

Figure 6.2. Person abilities (θ) estimated by maximum likelihood estimation. 30

Figure 6.3. Person abilities (θ) estimated by expected a priori methods. 30

List of Equations

Equation 1. Item Difficulty Index from CTT Analyses (1) 22
Equation 2. Standard Error of Measurement (SEM) from CTT Analyses (2) 26
Equation 3. Two-Parameter (2PLS) Model (3) 27
Equation 4. Graded Response Model (GRM) (4) 27
Equation 5. Conditional Standard Error of Measurement (CSEM) Given Person Ability (5) 28

Executive Summary

The web-based Knowledge for Teaching Elementary Fractions test, designed to measure mathematical knowledge for teaching (MKT) in the domain of fractions at the elementary level, was administered to a sample of 277 elementary educators, including teachers, administrators, and instructional support personnel, in fall 2016, 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 Knowledge for Teaching Elementary Fractions test is to produce ability estimates that can be used to investigate baseline equivalence of groups of educators in four treatment conditions, to serve as a covariate in models estimating the effect of the intervention on MKT, as well as to investigate MKT as a potential moderator of the effect of the program on teachers and students. In the present report, we discuss the development of the test, our exploration of options for scoring and data modeling, and decisions made to support optimal scoring and data-modeling procedures. We also report on the results of data modeling, including analyses of dimensionality, scale reliability estimates, item difficulty estimates, test information, and the distribution of educator ability estimates.

Description of the Test

The test's content is designed to align with the intersection of the Common Core State Standards for Mathematics and an intervention involving lesson study with a fractions resource toolkit (Lewis & Perry, 2017).

The full test form contained a combination of selected-response and constructed-response items, including fill-in-the-blank, short answer, and extended response questions. Most of the extended-response questions were designed for qualitative, categorical coding. Those items are excluded from the present analyses. The part of the test form designed for quantitative scoring contains 19 items, prompting up to 30 individual responses from the test taker. Twenty-five of the 30 responses use a selected-response format (including two yes/no responses), and the remaining five a constructed-response (fill-in-the-blank) format.

Sample and Setting

The test was administered to with a sample of 277 elementary educators in six U.S. states in fall 2016. Eleven of these responded to less than 75% of the items and were dropped from analysis, leaving an analytic sample of 266 educators for the present report.

A single test form was used for all subjects in the sample. The subjects were participating in a large-scale randomized controlled trial of lesson study with a fractions resource toolkit. The tests were administered as a web-based questionnaire using Qualtrics software and scored by research-project staff at Florida State University.

Results

Item Diagnostics and Scoring

Item diagnostics and calibration accounting resulted in collapse of the 30 individual responses (or nonresponses) into a total of 18 independent items. After one item was removed because of poor psychometric outcomes, the remaining 29 were included in the final 18-item scale.

Initial screening of the items used an approach based on classical test theory (CTT). The median p-values for the 18 items in the final scale was .60, the minimum value was .12, and the maximum value was .96, suggesting a broad range of difficulty among items on the test. The median item-rest correlation coefficient was .36, the minimum value was .22, and the maximum value was .48, suggesting that the items in the final scale had adequate 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 Knowledge for Teaching Elementary Fractions test data.

Item Response Theory 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 two-parameter logistic model and a graded response model (GRM) based on item-response theory (IRT). The models were run by means of flexMIRT (version 3.5) software (Cai, 2017). Findings from IRT analyses indicated that the item discrimination estimates ranged from 0.77 to 1.77 (M = 1.14, SD = 0.30).

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, but for computational reasons, it cannot provide person ability estimates for respondents who have perfect or zero test scores (de Ayala, 2009). To help estimate these extreme cases, we used an *expected a posteriori* (EAP) estimator.

Reliability and Test Information

By means of a CTT approach, coefficient α and standard error of measurement (SEM) were calculated to be .76 and 2.32, respectively. In addition, 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 (θ) was approximately 0.00. The person-ability estimate was associated with higher test information and lower CSEM for the person ability estimates between –2.00 and 2.00 on the θ scale and was associated with lower test information and higher CSEM for the person-ability estimates greater than 2 or less than –2 on the θ scale.

Distribution of Educator Ability Scores

Using an EAP technique, we found that the distribution of student ability (θ) scores for the educator in the present sample does not appear to differ from a normal distribution. By the EAP method, the θ estimates for the educators in the sample ranged from –2.86 to 2.35 (M = 0.00, SD = 0.90). The skewness and the kurtosis statistics for the sample distribution were 0.15 and –0.34, respectively.

Discussion and Conclusions

In summary, we found that the Knowledge for Teaching Elementary Fractions test 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, although further validation will be necessary to determine whether the test is well suited for its intended use. Evaluation of the structural validity of the resulting 18-item scale supports the assertion that the Knowledge for Teaching Elementary Fractions test meets or exceeds common standards for educational and psychological measurement for its stated purpose.

1. Introduction

The present report includes the scoring and data modeling of the Knowledge for Teaching Elementary Fractions test. The items on this test that comprise the final score were designed to measure content knowledge and specialized content knowledge (Ball, Thames, & Phelps, 2008) on the topic of fractions. Correct responses to items require teachers to understand related ideas such as referent unit, partitioning and iterating, identifying points on a number line corresponding to rational numbers, computation involving fractions, and representing word-problem scenarios involving fractions and operations on fractions with equations and expressions. The collections of items on the test are not designed to create subscales. Rather, the test is designed to measure a single (albeit broad) construct: mathematical knowledge for teaching elementary-level fractions concepts.

All the items on this test were borrowed or adapted from other sources, including the Diagnostic Teacher Assessment in Math and Science project (DTAMS; Saderholm, Ronau, Brown, & Collins, 2010), Learning Mathematics for Teaching project (LMT; Hill, Schilling, & Ball, 2004; LMT, 2004), Numeracy Development Projects (Ward & Thomas, 2015), and other publications (Beckmann, 2005; Newton, 2008; Norton & McCloskey, 2008; Schifter, 1998; Zhou, Peverly, & Xin, 2006).

A previous version of the test was used in a randomized trial investigating the impact of lesson study with fractions resource toolkits on teachers and students (Lewis & Perry, 2017). The previous version of the Knowledge for Teaching Elementary Fractions test detected a significant difference between teachers in a treatment condition and those in a control condition (Lewis & Perry, 2017). The version of the test used for the present sample was used as a baseline measure of fractions knowledge for teachers in a subsequent study involving a larger sample.

1.1. Description of the Sample

The present report focuses on the version of the Knowledge for Teaching Elementary Fractions test that was administered to a group of 277 educators in fall 2016. These educators represented six states in the U.S. Characteristics of the individuals in the sample are provided in Table 1.1. Approximately 81% of the sample were regular classroom teachers, the majority of whom were teaching third (42%), fourth (33%), or fifth (14%) grade. The average years of teaching experience among teachers in the sample was 12.8.

Table 1.1. Characteristics of Teachers in the Fall 2016 Field-Test Sample (N = 266)

Characteristic	Total (Proportion)
Primary teaching role	
Regular classroom[a]	215 (.811)
Varying exceptionalities[b]	15 (.056)
English language learners	2 (.008)
Other[c]	33 (.125)
Departmentalization	
Teaches all subjects	175 (.660)
Teaches only mathematics	79 (.298)
Does not teach mathematics	11 (.042)
Grade level primarily taught	
Kindergarten	2 (.008)
Grade 1	4 (.015)
Grade 2	14 (.053)
Grade 3	111 (.417)
Grade 4	87 (.327)
Grade 5	38 (.143)
Grade 6	6 (.023)
Grade 7	2 (.008)
Grade 8	1 (.004)
Highest degree earned	
No degree[d]	1 (.004)
Bachelor's degree	135 (.508)
Master's degree	112 (.421)
Specialist degree	18 (.068)
Areas of certification	
Elementary Education	242 (.910)
PreK/Primary Education	36 (.135)
Middle Grades Mathematics	20 (.075)
Secondary Mathematics	4 (.015)
ESOL/Bilingual/Dual-language	110 (.414)
Varying Exceptionalities[b]	72 (.271)
State	
Florida	176 (.662)
Illinois	33 (.124)
California	32 (.120)
Colorado	8 (.030)
Indiana	3 (.011)
New York	14 (.053)
Years of teaching experience	12.8 ± 7.5

Note. Statistics are presented as frequency (percentage) for categorical variables and mean ± standard deviation for numerical variables.

[a]Regular classroom teachers teach core content but may have classrooms where gifted and talented students, students with disabilities, and/or English language learners are enrolled.

[b]Varying exceptionalities indicates specialized instruction for gifted and talented students and students with disabilities.

[c]Other includes teachers of noncore subject areas, math coaches, and administrators.

[d]One respondent selected "do not have a degree" and only responded to the questions about degree earned and years of teaching experience. This leaves the other demographics with one participant fewer than the full sample of 266.

1.2. Detailed Test Blueprint

Table 1.2 contains a detailed blueprint for the items on the Knowledge for Teaching Elementary Fractions test. Many of the items were borrowed from existing item banks, and the others were adapted from published sources. An account of the source of each item is provided in Appendix A

Table 1.2. Test Blueprint for the Fall 2016 Knowledge for Teaching Elementary Fractions Test

Item description	Original #	Recoded #	Final scale #
Is possible as a fraction	1a	1	
Teacher action to respond to Anna	1b		
Number line point best representing	2	2	1*
Student representations of	3a	3	2*
Student representations of	3b	3	2*
Student representations of	3c	3	2*
Student representations of	3d	3	2*
Point closest to	4	4	3*
How number line can help students understand fractions	5		
Things students should understand about	6		
Relationship between numerator and denominator in	7	5	4*
Steve- fiction is more than Andrew fiction. Correct?	8a	6	5*
Why/why not is Steve not correct?	8b		
	9	7	6*
	10a	8	7*
	10b	8	7*
	10c	8	7*
Given yards rope, with per rope, how many ropes?	11	9	8*
Student representations of	12	10	9*
Jim's proportion of program sessions taught	13	11	10*
Word problem for	14a	12	11*
Word problem for	14b	12	11*
Word problem for	14c	12	11*
Word problem for	14d	12	11*
Divide rectangular cakes equally among students	15	13	12*
	16	14	13*
Models to represent	17	15	14*
Connections- measurement and fractions	18		
Fractional part of square in triangle A	19	16	15*
Paper frog moving along a line	20	17	16*
What would students need to know to solve these problems	21		
Why important for students to answer how many -s in ?	22		
Similarities/differences bet fractions/whole numbers	23		
Word problem 3 divided by	24a	18	17*
Word problem 3 divided by	24b	18	17*
Word problem 3 divided by	24c	18	17*
Word problem 3 divided by	24d	18	17*
Comparing and	25	19	18*

Note. Italicized item descriptions correspond to items that do not contribute to the quantified test score. Item description = the description of an item that requires a response; original # = the original index number of each item; recoded # = the item index number after excluding qualitative items and forming polytomously scored items; final # = the item index number (with a * after the number to help differentiate from the recoded item index number) in the final scale.

2. Initial Item Review

The Knowledge for Teaching Elementary Fractions test consists of 25 numbered items that require assessed teachers to make a total of 38 responses, because items 1, 3, 8, 10, 14, and 24 require multiple responses. (See Appendixes A and B for specifics.) The 38 responses can therefore be split into two groups, of which the first consists of 30 responses that can be scored as correct or incorrect. These correspond to either selected-response or constructed-response (fill-in-the-blank) items.

The other eight responses, designed to be coded by descriptive categories, are intended to provide insight into teachers' thinking processes or perspective on teaching and learning fractions; these answers are not designed to be judged correct or incorrect. Because the present report is a quantitative investigation of the Knowledge for Teaching Elementary Fractions test, these eight items were dropped from data entry, leaving just 19 items in the recoded test. Table 1.2 presents the details of this recoding process.

During data entry, the 30 fraction-focused responses in the recoded test were scored dichotomously as correct or incorrect in accordance with the answer keys. Because some recoded items (i.e., item 3, 8, 12, 18) require multiple responses, we scored these items polytomously by summing the scores of their responses. The recoding was performed to address concerns about local dependence of responses within items, because we used item-response-theory models in scoring teachers' latent ability. During subsequent statistical analysis, we further adjusted the test by removing item 1 in the recoded test. The final version of the test therefore consisted of 18 items. We placed an asterisk after the tem numbers on the final test to avoid confusion with the item numbers on the recoded test. Table 1.2 shows the correspondence between the two numbering systems.

The changes to the test were not necessarily performed in the order they are reported here but were the result of an interactive, overlapping, and iterative process. For example, the decision to remove item 1 from the recoded test was informed by results of different analyses, such as those following classical test theory and exploratory factor analysis.

3. Data and Scoring

3.1. Data Entry and Verification Procedures

The Knowledge for Teaching Elementary Fractions test was administered as an online survey using Qualtrics software. Response data were exported from Qualtrics to a flat file and manipulated by means of SPSS and Excel software.

Selected-response items were scored according to the predetermined scoring guide provided in Appendix A. The responses to the constructed-response items were reviewed during an adjudication meeting with a committee comprising experts in mathematics, mathematics education, and mathematics teacher education. The adjudication committee reviewed the full set of unique responses to determine the set of correct responses, which are provided in Appendix A.

Teachers were given the freedom to skip items, exit the test at any time, and retake the test at any time during the testing window. This freedom in testing conditions sometimes created multiple submissions for participants. When participants submitted multiple responses for a given item, their final response was taken to be that with the latest date.

3.2. Item Scoring

A total of 277 teachers took the Knowledge for Teaching Elementary Fractions test, but not every teacher gave complete responses. The decision was therefore made to exclude teachers who had a response rate lower than 75%. That is, teachers were removed from the set who had eight or more missing responses out of 30. Although the decision to 75% as the cut-off point is arbitrary, it does seem to align with a pattern in the excluded cases. Specifically, the excluded teachers showed more missing responses in the second half of the test, a pattern that seemed to imply a lack of motivation to complete the test. They were allowed to stop in the middle of the test and continue the test at a later time. Table 3.1 shows the frequency of teachers' missing response(s) in the sample.

After the eight responses not intended to be used in the test score were excluded, the recoded test consisted of 19 items, resulting in a possible 30 responses from teachers. These responses were scored according to answer keys provided by test developers. The answer key and scoring criteria are provided in Appendix B.

Some items prompted multiple responses from the same item stem. For example, item 3 of the original test requires four responses from teachers, and teachers' scores on item 3 are represented by a polytomous variable defined as the sum of four dichotomous variables, corresponding to the four responses (see Table 1.2). Generating polytomously scored items is necessary for addressing the local dependence issue when using item response theory to estimate teachers' latent ability.

After the data from the recoded test was analyzed by means of statistical models, consistent evidence indicated that the recoded test should be further revised. Specifically, five pieces of evidence suggested removing item 1. First, the interitem correlation coefficients between item 1 and the rest of the items were low, ranging from –0.12 to 0.12. Second, the corrected item-total (i.e., item-rest) correlation coefficient (0.10) for item 1 was low and below the commonly suggested minimum of .20. Third, on the basis of the calculation of correct response rate, the estimated item difficulty for item 1 was 0.95, suggesting that item 1 was a very easy item for the teachers tested. Fourth, coefficient α (Cronbach, 1951) increased (to .76) when item 1 was removed from the scale. Last, results from exploratory factor analysis suggested that, unlike other items that loaded heavily on one latent factor, item 1 tended to

load on a second latent factor, on which other items showed small loadings. Given these results, we decided to remove item 1 from the final scale. This test revision resulted in a final scale consisting of 18 items (see Table 1.2). The remainder of the present report focuses on results from analysis of the final-scale test.

Table 3.1. Missing Response Frequency in the Sample

No. of Missing response(s)	Frequency	%	Cumulative %
0	239	86.28	86.28
1	16	5.78	92.06
2	7	2.53	94.58
3	2	0.72	95.31
5	2	0.72	96.03
8	1†	0.36	96.39
10	1†	0.36	96.75
13	1†	0.36	97.11
14	2†	0.72	97.83
16	1†	0.36	98.19
17	1†	0.36	98.56
18	1†	0.36	98.92
20	2†	0.72	99.64
25	1†	0.36	100.00
Total	277	100.00	

Note.
†Teachers excluded from the analysis. # of Missing response(s) = the number of missing response(s) for a given teacher in the sample; frequency = the number of teachers with a given number of missing response(s); % = the percentage of teachers who had given numbers of missing response(s); cumulative % = cumulative percentage of teachers who had given numbers of missing response(s).

4. Dimensionality Analysis

The data consisted of dichotomously and polytomously scored items. Because of the polychoric correlation of the data, we conducted exploratory factory analysis using Mplus 7.0 (Muthén & Muthén, 1998-2012) to investigate its dimensionality. Table 4.1 shows the eigenvalues and corresponding variation explained by each component. These eigenvalues are also presented in the scree plot in Figure 4.1. The largest eigenvalue was 6.02, and the first component explained 33% of the variation.

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

Component	Eigenvalue	% of variation explained
1	6.02	33.44
2	1.37	7.61
3	1.27	7.06
4	1.22	6.78
5	1.09	6.06
6	0.89	4.94
7	0.85	4.72
8	0.77	4.28
9	0.75	4.17
10	0.70	3.89
11	0.61	3.39
12	0.60	3.33
13	0.55	3.06
14	0.42	2.33
15	0.32	1.78
16	0.25	1.39
17	0.21	1.17
18	0.12	0.67

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

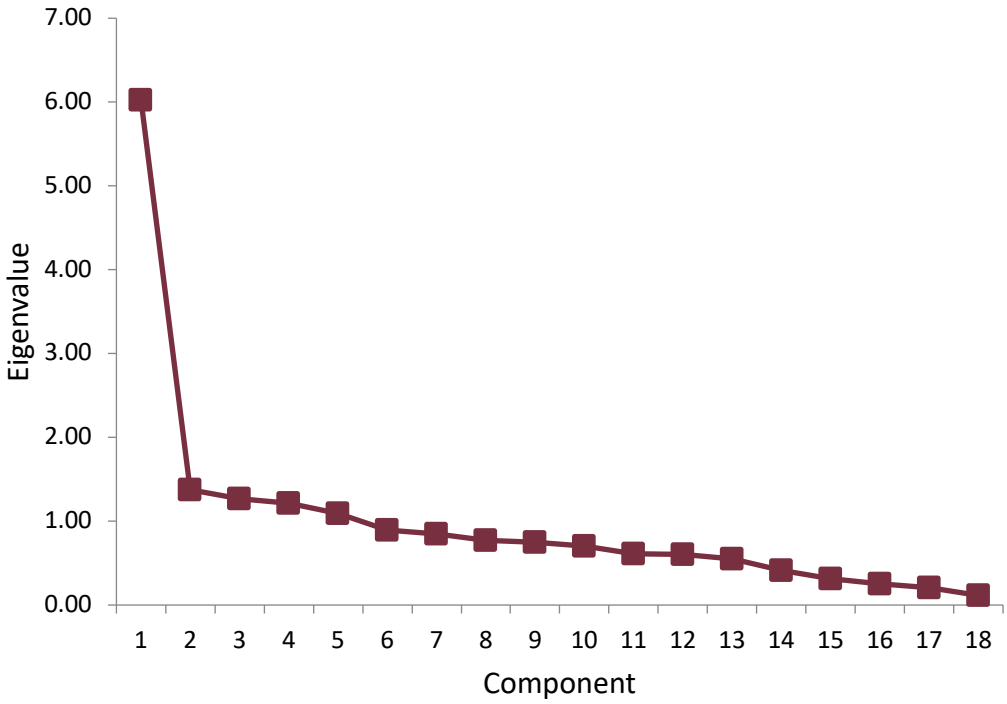

Figure 4.1. Scree plot of eigenvalues estimated from Mplus.

In addition, we performed parallel analysis to examine the dimensionality of the data further, using the *psych* (Revelle, 2017) package in R 3.4.0 (R Core Team, 2012). The results supported unidimensionality, so the explanatory factor analysis and the parallel analysis results seemed to indicate a single, dominant factor in the data.

5. Classical Testing Theory (CTT) Analyses

5.1. Distribution of the Observed Test Score

We conducted the CTT analyses using SPSS 22.0 (IBM corp., 2013). Figure 5.1 displays the distribution of observed sum scores in the final-scale format. The mean of the observed test scores was 18.72, and the standard deviation was 4.73. The median was 19.00, the mode was 20.00, the skewness was –0.14, and the kurtosis was –0.34. Note that although the final-scale format had only 18 items, the observed test score ranged from 2.00 to 29.00, because the items 2*, 7*, 11*, 17* were coded into polytomous items.

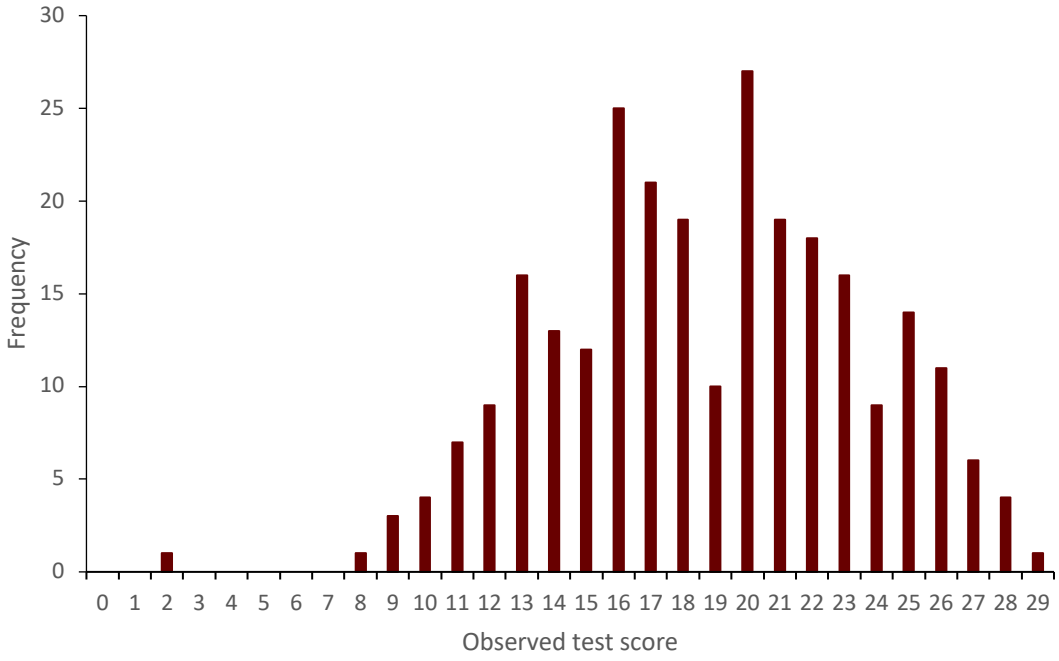

Figure 5.1. Distribution of the observed test scores in the final-scale format.

5.2. Item Difficulty & Discrimination

The item difficulty and item discrimination of the final-scale items were first estimated by means of CTT-based analyses. Equation 1 shows the formula used to calculate the CTT-based difficulty index,

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

where p is the symbol of the item difficulty index (McDonald, 1999). For dichotomous items, the difficulty index calculation is equivalent to the proportion of correct answers.

Table 5.1 shows the mean, standard deviation, item difficulty, and item discrimination estimates of each final-scale item. The item difficulty varied from a maximum of .12 (item 9*) to a minimum of .96 (item 2*). The mean of the item difficulty estimates was .59, standard deviation .21. The skewness statistic of the item difficulty estimates in the test was –0.24, and the kurtosis statistic was 0.18. To investigate item discrimination, we calculated the item-rest correlation coefficients (i.e., corrected item-total correlation

coefficients) for each of the items. The item discrimination estimates varied from a minimum of .22 (item 2*) to a maximum of .48 (item 6*). The discrimination estimates for all the items were greater than .20. The mean of the item discrimination estimates was .36, standard deviation 0.07. The skewness statistic was –0.12, and the kurtosis statistic was –0.23.

Table 5.1. Item Difficulty and Discrimination from CTT Analyses

Final-scale item #	*M*	*SD*	*p*	Item-rest *r*
1*	0.48	0.50	.48	.41
2*	3.82	0.49	.96	.22
3*	0.64	0.48	.64	.42
4*	0.87	0.34	.87	.31
5*	0.74	0.44	.74	.37
6*	0.42	0.49	.42	.48
7*	2.37	0.82	.79	.35
8*	0.80	0.40	.80	.36
9*	0.12	0.33	.12	.33
10*	0.45	0.50	.45	.46
11*	2.47	0.96	.62	.36
12*	0.59	0.49	.59	.32
13*	0.46	0.50	.46	.34
14*	0.51	0.50	.51	.29
15*	0.74	0.44	.74	.36
16*	0.45	0.50	.45	.44
17*	2.43	1.22	.61	.43
18*	0.35	0.48	.35	.28

Note. Final-Scale Item # = forming polytomously scored items and removing a problematic item (we differentiated recoded item index and final-scale item index by adding an asterisk to 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. Figures 5.2 and 5.3 display the item difficulty and item discrimination, respectively.

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

p-value	Number of items
.90–1.00	1
.80–.89	2
.70–.79	3
.60–.69	3
.50–.59	2
.40–.49	5
.30–.39	1
.20–.29	0
.10–.19	1
.00–.09	0
Mean	.59
Standard Deviation	.21
Minimum	.96
Maximum	.12

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
.80–1.00	0
.60– 79	0
.40–.59	6
.20–.39	12
.00–.19	0
Mean	.36
Standard Deviation	.07
Minimum	.22
Maximum	.48

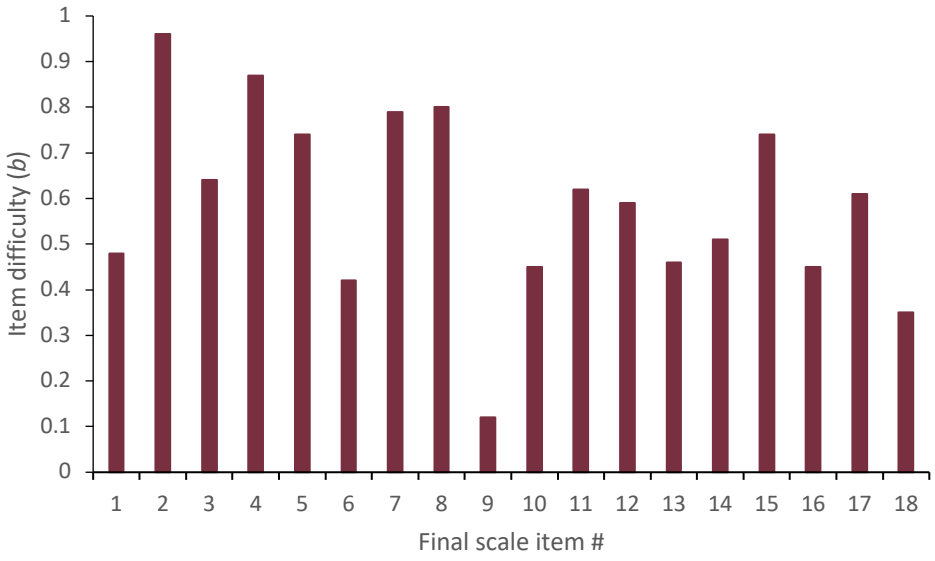

Figure 5.2. Item difficulty estimate (b) of each final-scale item.

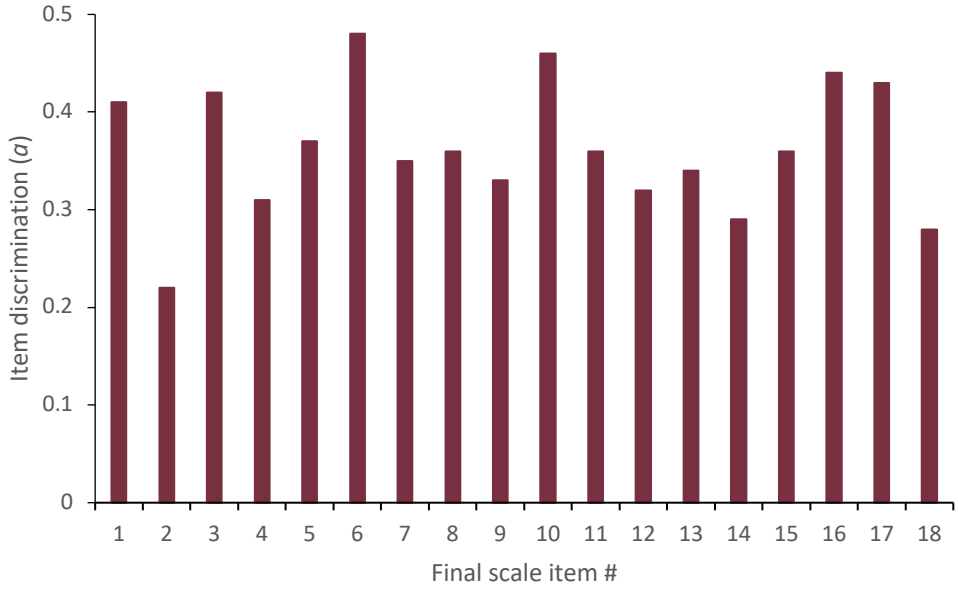

Figure 5.3. Item discrimination estimate (a) of each final-scale item.

5.3. Coefficient α and Standard Error of Measurement

We calculated coefficient α (Cronbach, 1951) as one way to estimate the test reliability. The estimated coefficient α of the test was 0.76. We subsequently calculated the standard error of measurement (SEM) of the test. SPSS output indicated that the scale variance was 22.34. On the basis of Equation 2, SEM was calculated to be 2.32.

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

where σ^2 is the test variance, and ρ_{XX} is the coefficient α of the test.

6. Item-Response Theory Analyses

6.1. Model Description

We conducted item-response theory (IRT) analyses using the software flexMIRT 3.5 (Cai, 2017). For the dichotomous items (1*, 3*, 4*, 5*, 6*, 8*, 9*, 10*, 12*, 13*, 14*, 15*, 16*, and 18*), a two-parameter (2PL) model was used. For the polytomous items (2*, 7*, 11*, and 17*), a graded response model (GRM) was used.

Results of FlexMIRT indicated that successful convergence was reached in the computation, and the value of -2loglikelihood was 6317.30. The formulas of the 2PL model and GRM, based on the parameterization of De Ayala (2009), are provided in Equations 3 and 4.

The formula used for the 2PL model was

$$P_j(\Theta) = \frac{\exp\left[a_j(\Theta - b_j)\right]}{1+\exp\left[a_j(\Theta - b_j)\right]}, \quad (3)$$

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, and Θ is the person ability.

The formula used for the GRM model was

$$P_{jk}(\Theta) = \frac{\exp[a_j(\Theta - b_{jk})]}{1+\exp[a_j(\Theta - b_{jk})]}, \quad (4)$$

where a_j is the discrimination index of item *j* (*j = 1, 2,…,J*), P_{jk} is the probability of category *k* or higher, *k* $\in \{0, 1, 2, \ldots, k\}$, Θ is the person ability, and b_{jk} is category threshold.

6.2. Item Difficulty and Discrimination

Tables 6.1 and 6.2 present parameter estimates of the 2PL- and GRM-modeled items, respectively. The discrimination estimates for the 18 items ranged from 0.77 (item 14*) to 1.77 (item 6*), and 11 items (1*, 3*, 4*, 5*, 6*, 8*, 9*, 10*, 15*, 16*, and 17*) had values above 1.00. For all the 18 items, the mean of the item discrimination estimates was 1.14, standard deviation 0.30. The skewness statistic of the item discrimination estimates was 0.54, and the kurtosis statistic was –0.57. For the 14 items using 2PL models shown in Table 5, the item difficulty estimates ranged from a minimum of –1.78 (item 4*) to a maximum of 1.85 (item 9*). The mean of the item difficulty estimates for the 14 items using 2PL models was –0.21, standard deviation 0.96. The skewness statistic of the item difficulty estimates for the 14 items using 2PL model was 0.33, and the kurtosis statistic was 0.35.

Table 6.1. Parameter Estimates and Standard Errors for Final-Scale Items Modeled Using a Two-Parameter Model

Final-scale item #	*a* (*SE*)	*b* (*SE*)
1*	1.20 (0.23)	0.07 (0.14)
3*	1.13 (0.23)	–0.63 (0.18)
4*	1.39 (0.45)	–1.78 (0.42)
5*	1.26 (0.26)	–1.09 (0.22)
6*	1.77 (0.32)	0.26 (0.12)
8*	1.32 (0.34)	–1.38 (0.28)
9*	1.44 (0.35)	1.85 (0.32)
10*	1.65 (0.30)	0.16 (0.12)
12*	0.91 (0.21)	–0.46 (0.20)
13*	0.89 (0.20)	0.19 (0.18)
14*	0.77 (0.19)	–0.05 (0.20)
15*	1.17 (0.30)	–1.15 (0.25)
16*	1.32 (0.25)	0.20 (0.14)
18*	0.78 (0.19)	0.87 (0.25)

Note. Final-Scale Item # = the newly generated item number after formation of polytomously scored items and removal of a problematic item (asterisks follows item numbers used in the final scale); *a* = item discrimination index; *b* = item difficulty index; *SE* = standard error.

Table 6.2. Parameter Estimates and Standard Errors for Final-Scale Items Modeled by Means of a Graded-Response Model

Final-scale item #	*a* (*SE*)	b_1 (*SE*)	b_2 (*SE*)	b_3 (*SE*)	b_4(*SE*)
2*	0.85 (0.26)	–5.63 (1.66)	–4.59 (1.31)	–2.40 (0.64)	
7*	0.82 (0.19)	–4.96 (1.08)	–2.16 (0.46)	–0.35 (0.20)	
11*	0.85 (0.18)	–3.71 (0.75)	–2.56 (0.55)	–0.27 (0.19)	2.89 (0.56)
17*	1.02 (0.17)	–2.62 (0.45)	–1.55 (0.29)	–0.03 (0.16)	1.43 (0.26)

Note. Final-Scale Item # = the newly generated item number after forming polytomously scored items and removing a problematic item (asterisks follows item numbers used in the final scale); *a* = item discrimination index; b_{jk} $(j = 1,2, \ldots ,18, k = 0, 1, 2, 3, 4)$ = category threshold; *SE* = standard error.

6.3. Test Information and Estimated Person Ability

Figure 6.1 displays the test information curve and the conditional standard error of measurement (CSEM) for the test. Equation 5 shows the formula used in the CSEM calculation

$$CSEM(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (5)$$

where I is the test information function for a given person ability, and θ is the person ability (De Ayala, 2009).

According to the relationship between test information and CSEM, a person ability (θ) estimate around the value of 0.00 was associated with the highest test information and the lowest CSEM. In addition, the CSEM curve in Figure 6.1 suggested that the person-ability estimates were related to lower CSEM (i.e.,

more accurate estimation of person ability) when it ranged between –1.0 and 1.0; the curve also suggested that person ability estimates were related to higher CSEM (i.e., less accurate estimation of person ability) when it was larger than 2 or less than –2.

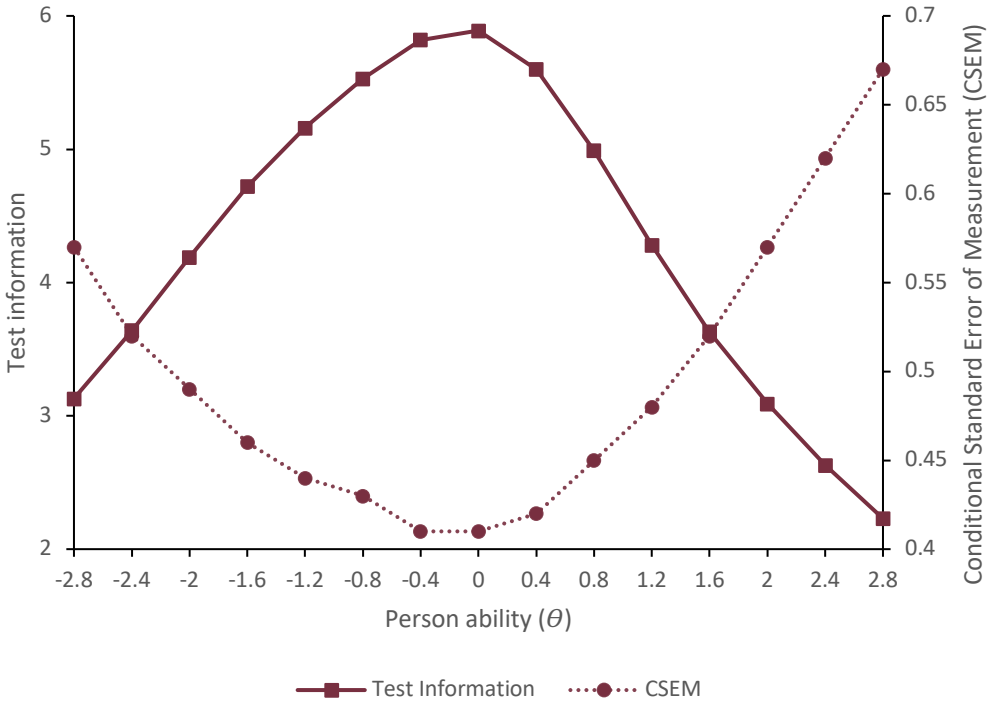

Figure 6.1. Test information curve and conditional standard error of measurement for the final-scale items.

Figure 6.2 presents the person ability estimates using the maximum likelihood estimation (MLE) method. For individuals who get perfect or zero scores, the MLE ability estimates are not available. In this sample, no teacher got a zero score, and just one got a perfect score. When person ability is estimated by MLE, the minimum and the maximum likelihood scores were set as -7 and 7, respectively, in the flexMIRT software.

We also used expected *a posteriori* (EAP) method to estimate person ability. Figure 6.3 presents the distribution of person ability from EAP. The person-ability scores ranged from –2.86 to 2.35. The mean and standard deviation of the EAP estimates were 0.00017 and 0.90, respectively. The skewness and the kurtosis of the person ability were 0.15 and –0.34, respectively.

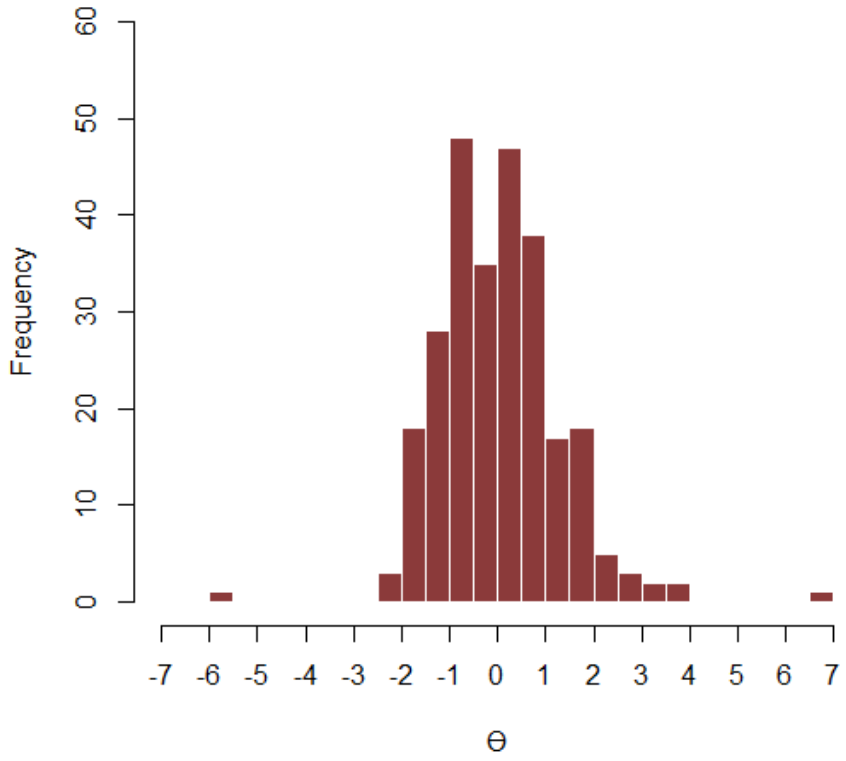

Figure 6.2. Person abilities (θ) estimated by maximum likelihood estimation.

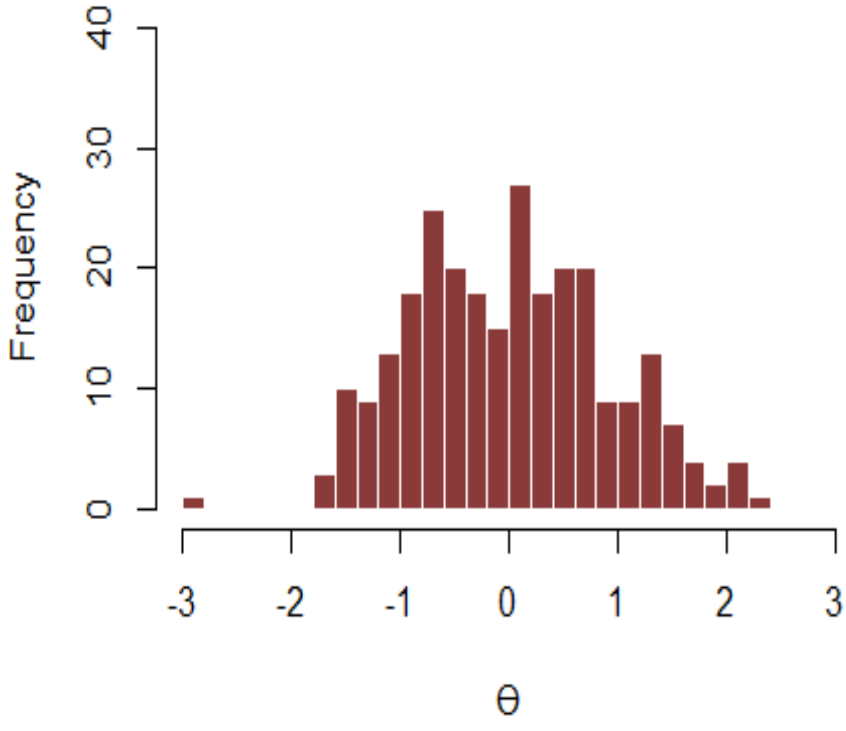

Figure 6.3. Person abilities (θ) estimated by expected a priori methods.

7. Discussion

Here we report findings from a field test of the Knowledge for Teaching Elementary Fractions test during fall 2016. This psychometric report provides several important contributions to the validation of the test. We discuss some of those results below, organized according to a three-part framework for test validation provided by Flake, Pek, and Hehman (2017).

7.1. Substantive Validity

All the items on the test were copied or adapted from other published sources. Each of those sources was subject to expert and/or peer review. In addition, the items were reviewed by content experts who are part of the senior personnel or the advisory board for the randomized controlled trial. The items were found to be accurate with respect to content and aligned to the types of MKT relevant to teaching fractions at the elementary level in accordance with the Common Core State Standards for mathematics (NGACBP & CCSSO, 2010).

The test was not designed or organized according to subcategories within the domain of fractions. Considering the finding that the test measures a unidimensional construct, subcategories may not be necessary, but they may provide additional description and support for the interpretation of scores. For example, the items could be sorted according to categories such as referent unit, partitioning and iterating, and relative magnitude of fractions. It could also be split according to content and pedagogical content knowledge or by domains within more specific theoretical frameworks for MKT (Ball et al., 2008). For example, interpretation of linear representations of fractions or identification of points on the number line corresponding to fractions might be considered either common content knowledge or specialized content knowledge.

7.2. Structural Validity

7.2.1. Unidimensionality

Exploratory factor analysis and parallel analysis both indicated a single, dominant factor in the data. This result suggests the Knowledge for Teaching Elementary Fractions test may be measuring a single, MKT-related latent construct. Alternatively, it might suggest that the MKT can be considered unidimensional. More research is needed to determine whether the theorized facets of MKT can be identified from empirical data generated by measurement instruments designed to distinguish among the various facets.

7.2.2. Level of Difficulty for the Intended Population

The difficulty of the test aligned well with the ability level of the educators in the sample. Moreover, the distribution of scores appear to be reasonably close to a normal distribution, which is how we might expect the population of educator abilities to be distributed. No participant received a zero score on the final test scale, and only one participant received a perfect score. On the basis of the CTT results, the item difficulty estimates ranged from .12 to .96. The mean was .59 with a standard, 0.21. The item discrimination estimates ranged from .22 to .48. The mean was .36 with a standard deviation of 0.07.

7.2.3. Test Information

According to the relationship between test information and CSEM, a person ability estimate (θ) around the value of 0.00 was associated with the highest test information and the lowest CSEM. Person-ability estimates were related to lower CSEM (i.e., more accurate estimation of person ability) when person ability ranged between –1.0 and 1.0, a result that aligns with the ability estimates for 71.43% of the educators in the sample (based on the EAP θ estimation). Person-ability estimates were related to higher CSEM (i.e., less accurate estimation of person ability) when it was larger than 2 or less than –2. Note, however, that those extreme person-ability estimates were observed in only six cases, comprising 2.26% of the total sample (on the basis of the EAP θ estimation).

7.3. External Validity

The Knowledge for Teaching Elementary Fractions test will be used as a covariate in the models designed to estimate the effect of the intervention on educators' MKT. The MKT posttest is identical to the present test with the exception of one item. As a result, we anticipate the teacher scores to be a strong predictor of their posttest scores. If they are ultimately found to not be a strong predictor, the conditions under which the test is administered should be examined for flaws. In this case, which is not expected, follow up with participants through brief interviews might be advisable.

A previous version of the Knowledge for Teaching Elementary Fractions test was used in a previous randomized trial (Lewis & Perry, 2017). Using CTT-based scoring methods, the previous version of the test detected a significant difference in teacher performance among the teachers in the treatment and control groups. We do not yet know how the IRT-based scoring method might affect the ability of the test to detect a treatment effect, but IRT-based methods might reasonably be expected to increase the ability of the Knowledge for Teaching Elementary Fractions test to detect a treatment effect. The results of those analyses are not available at the time of the writing the present report. Likewise, whether the scores on the Knowledge for Teaching Elementary Fractions test will significantly predict student learning or moderate the effect of the intervention on student learning is not yet known.

7.4. Conclusions

On the basis of the sample of 266 educators from fall 2016, the Knowledge for Teaching Elementary Fractions test appears to measure a dominant factor, supporting unidimensionality in the data. Reliability, test information, and item-discrimination estimates appear to fit the intended purpose of the test, although further validation will be necessary to determine the extent to which the test is well-suited for its intended use. Evaluation of the structural validity of the resulting 18-item scale supports the assertion that the test meets or exceeds common standards for educational and psychological measurement for its stated purpose.

The overall difficulty of the test appears to align well with the intended population. One examinee received a perfect score, and the ability estimate for one examinee was extremely low. The person ability of the participant who received the perfect score cannot be estimated with the MLE estimator, but it can be estimated with the EAP estimator. The distribution of the person-ability estimates with the EAP estimator had a mean near zero and standard deviation of .90. As a result, the person-ability estimates resulting from the EAP estimation are recommended for use in the anticipated statistical models estimating the effect of the intervention on educator knowledge and the effect of educator knowledge on student learning.

References

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407.

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.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: The Guilford Press.

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.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematical knowledge for teaching. *The Elementary School Journal, 105*(1), 11–30.

IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

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.

LMT (Learning Mathematics for Teaching). (2004). Mathematical knowledge for teaching measures: Geometry content knowledge, number concepts and operations content knowledge, and patterns and algebra content knowledge. Ann Arbor, MI: Author.

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.

Newton, K. J. (2008). An extensive analysis of preservice elementary teachers' knowledge of fractions. *American Educational Research Journal, 45*(4), 1080–1110.

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

Norton, A. H., & McCloskey, A. V. (2008). Modeling students' mathematics using Steffe's fraction schemes. *Teaching Children Mathematics, 15*(1), 48–54.

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, IL. https://CRAN.R-project.org/package=psych Version = 1.7.8.

Saderholm, J., Ronau, R., Brown, E. T., & Collins, G. (2010). Validation of the Diagnostic Teacher Assessment of Mathematics and Science (DTAMS) instrument. *School Science and Mathematics, 110*(4), 180–192.

Schifter, D. (1998). Learning mathematics for teaching: From a teachers' seminar to the classroom. *Journal of Mathematics Teacher Education, 1*(1), 55–87.

Ward, J. & Thomas, G. (2015). Numeracy Development Project. Retrieved from http://www.nzmaths.co.nz/sites/default/files/Numeracy/FractionsScenarios.pdf

Zhou, Z., Peverly, S. T., & Xin, T. (2006). Knowing and teaching fractions: A cross-cultural study of American and Chinese mathematics teachers. *Contemporary Educational Psychology, 31*(4), 438–457.

Appendix A. Sources of Assessment Items

Item Number	Correct Response	Item Description	Item Original Source	Coded Qualitatively?
Q1A	1 (Yes)		Ward & Thomas, 2015	N
Q1B	–	Teacher action to respond to Anna		Y
Q2	D (4)	Number line point best representing	Saderholm, Ronau, Brown, & Collins, 2010	N
Q3A	1	Student representations of	Learning Mathematics for Teaching (LMT) [1]	N
Q3B	2	Student representations of		N
Q3C	1	Student representations of t		N
Q3D	1	Student representations of line		N
Q4	A (1)	Point closest to	LMT [2]	N
Q5	–	How number line can help students understand fractions	Mills College Lesson Study Group (MCLSG)	Y
Q6	–	Things students should understand about	MCLSG	Y
Q7	B (2)	Relationship between numerator and denominator in	Saderholm, Ronau, Brown, & Collins, 2010	N
Q8A[a]	No; Maybe; There is not enough information	Steve – fiction is more than Andrew fiction. Correct?	Ward & Thomas, 2015	N
Q8B	–	Why/ why not is Steve necessarily correct?	Ward & Thomas, 2015	Y
Q8C	–	Teacher action to respond to Steve	Ward & Thomas, 2015	Y
Q9[a]	75; 75 miles		Zhou, Peverly, & Xin, 2006	N
Q10A[a]	0		Newton, 2008	N
Q10B[a]	16		Newton, 2008	N
Q10C[a]	3; 90/30; 9/3		Newton, 2008	N

Q11[a]	2; 2.8; 2 with 2/3 left over	Given yards rope, with per rope, how many ropes?	Schifter, 1998	Y
Q12	E (5)	Student representations of	LMT [3]	N
Q13	C (3)	Jim's proportion of program sessions taught	LMT [4]	N
Q14A	2	Word problem for	LMT [5]	N
Q14B	2	Word problem for		N
Q14C	1	Word problem for		N
Q14D	1	Word problem for ½		N
Q15	B (2)	Divide rectangular cakes equally among students	LMT [6]	N
Q16	E (5)		LMT [7]	N
Q17	–	Line segment of	Beckmann, 2005	Y
Q18	C (3)	Models to represent	LMT [8]	N
Q19	--	Connections - measurement and fractions	MCLSG	Y
Q20	C (3)	Fractional part of square is triangle A	LMT [9]	N
Q21	C (3)	Paper frog moving along a line	LMT [10]	N
Q22A	–	Given draw the whole	Norton & McCloskey, 2008	Y
Q22B	–	What would students need to know to solve these problems?	MCLSG	Y
Q23	–	Why important for st to answer "how many in ?"	MCLSG	Y
Q24	–	Similarities/ differences bet fractions/ whole numbers	MCLSG	Y
Q25A	2	Word problem	LMT [11]	N
Q25B	1	Word problem		N
Q25C	2	Word problem		N
Q25D	1	Word problem		N
Q26	B (2)	Comparing	LMT [12]	N
Note.				

[a]These items were formatted as constructed-response. The set of responses listed in the Correct Response column comprise the full set of responses observed in the data and determined to be mathematically valid and correct responses to the item prompt by the adjudication committee.

[1] Elementary Number Concepts & Operations, Content Knowledge, 2001B-1
[2] Elementary Number Concepts & Operations, Content Knowledge, 2001A-16
[3] Rational Number, Form B-1
[4] Elementary Number Concepts & Operations, Content Knowledge, 2001B-3
[5] Rational Number, Form B-9
[6] Elementary Number Concepts & Operations, Knowledge of Content and Students, 2001A-13
[7] Rational Number, Form A-6
[8] Elementary Number Concepts & Operations, Content Knowledge, 2001B-17
[9] Elementary Number Concepts & Operations, Content Knowledge, 2001B-5
[10] Rational Number, Form A-4
[11] Rational Number, Form A-10
[12] Rational Number, Form B-6

Appendix B. Knowledge for Teaching Elementary Fractions Test

The test items have been redacted from this report because we do not have the right to publish copywritten test items. Contact the lead author (rschoen@lsi.fsu.edu) for more information.