Psychometric Report on the Knowledge for Teaching Elementary Fractions Test Administered to Elementary Educators in Six States in Spring 2017

Robert C. Schoen
Xiaotong Yang
Insu Paek

AUGUST 2018

Research Report No. 2018-13

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., & Paek, I. (2018). *Psychometric report on the Knowledge for Teaching Elementary Fractions test administered to elementary educators in six states in spring 2017* (Research Report No. 2018-13). Tallahassee, FL: Learning Systems Institute, Florida State University. https://doi.org/10.17125/fsu.1537295574

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 Spring 2017

Research Report No. 2018-13

Robert C. Schoen

Xiaotong Yang

Insu Paek

August 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 performed the dimensionality analysis and item-response theory model calibration and wrote much of the 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.

In their roles as members of the project advisory board, Akihiko Takahashi, Tad Watanabe, Geoffrey Saxe, and Phil Daro reviewed the items on the initial draft test form and provided valuable feedback. 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 (Saderholm, Ronau, Brown, & Collins, 2010), the Learning Mathematics for Teaching project (Hill, Schilling, & Ball, 2004; LMT, 2004), Numeracy Development Projects (Ward & Thomas, 2015), the Knowledge for Teaching Elementary Mathematics project (Schoen, Bray, Wolfe, Tazaz, & Nielsen, 2017), the Mills College Lesson Study Group (https://lessonresearch.net), 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 iv

Executive Summary xi

Purpose Statement xi

Description of the Test xi

Sample and Setting xi

Dimensionality xii

Item Diagnostics and Scoring xii

Item-Response Theory Data Modeling xii

Reliability and Test Information xii

Distribution of Educator Ability Scores xii

Test-Score Equating xiii

Predictive Validity xiii

Discussion and Conclusions xiii

1. Introduction 1

1.1. Description of the Sample 1

1.2. Detailed Test Blueprint 3

2. Initial Item Review 5

3. Data and Scoring 6

3.1. Data Entry and Verification Procedures 6

3.2 Missing Data 6

3.3. Item Scoring 7

4. Dimensionality Analysis 8

5. Classical Testing Theory (CTT) Analyses 9

5.1. Distribution of the Observed Test Score 9

5.2. Item Difficulty & Discrimination 9

5.3. Coefficient α and Standard Error of Measurement 13

6. Item Response Theory (IRT) Analyses 14

6.1. Model Description 14

6.2. Item Difficulty and Discrimination 14

6.3. Test Information and Estimated Person Ability 15
6.4. Equating Posttest and Pretest Scores 18
6.4.1. Equating Process 18
6.4.2. Item Difficulty and Discrimination 18
6.4.3. Test Information and Estimated Person Ability 19
6.4.4. Predictive Validity 20
7. Discussion 22
7.1. Substantive Validity 22
7.2. Structural Validity 22
7.2.1. Unidimensionality 22
7.2.2. Level of Difficulty for the Intended Population 22
7.2.3. Test Information 22
7.3. External Validity 23
7.4. Conclusions 23
References 25

List of Appendices

Appendix A. Sources of Assessment Items 27

Appendix B. Knowledge for Teaching Elementary Fractions Test.....28

List of Tables

Table 1.1. Characteristics of Teachers in the Spring 2017 Field-Test Sample (N = 241) 2

Table 1.2. Test Blueprint for the Spring 2017 Knowledge for Teaching Elementary Fractions Test 4

Table 3.1. Missing Response Frequency in the Sample 6

Table 5.1. Item Difficulty and Discrimination Estimates Based on CTT Analyses 10

Table 5.2. Distribution of CTT-based Item-Difficulty (p-values) Estimates for Items Used in the Final Scale 11

Table 5.3. Distribution of CTT-based Item Discrimination (Item-Rest r) Point Estimates for Items Used in the Final Scale 11

Table 6.1. Parameter Estimates and Standard Errors for Final-Scale Items Modeled with Two-Paramater Logistic Methods 15

Table 6.2. Parameter Estimates and Standard Errors for Final-Scale Items Modeled with the Graded Response Model 15

Table 6.3. Item-Parameter Values of Equated Postintervention-Test Final-Scale Items Modeled from Two-Parameter Logistic Methods 19

Table 6.4. Item Parameter Values of Equated Posttest Final-Scale Items Modeled Using the Graded Response Model 19

List of Figures

Figure 4.1. Parallel analysis scree plot. PC, principal components.....8

Figure 5.1. Distribution of the observed test scores in the final-scale format. 9

Figure 5.2. Item difficulty estimate of each final-scale item.....12

Figure 5.3. Item discrimination estimate of each final-scale item. 12

Figure 6.1. Test information curve and CSEM for the final-scale items. 16

Figure 6.2. Sample distribution of person abilities (θ) estimated by maximum-likelihood estimation. 17

Figure 6.3. Sample distribution of person abilities (θ) estimated by expected a priori method. 17

Figure 6.4. Test information curve and conditional standard error of measurement for the equated posttest final scale.....20

Figure 6.5. Equated Person abilities (Θ) estimated by maximum likelihood estimation.....21

Figure 6.6. Equated Person abilities (Θ) estimated by the estimated a priori method.....21

List of Equations

Equation 1. Item difficulty index from CTT analyses (1) 9

Equation 2. Standard error of measurement (SEM) from CTT analyses (2) 13

Equation 3. Two-parameter logistic (2PL) item response theory (IRT) model (3) 14

Equation 4. Graded response model (GRM) (4) 14

Equation 5. Conditional standard error of measurement (CSEM) given person ability (5) 15

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 in spring 2017 to a sample of 241 elementary educators, including teachers, administrators, and instructional support personnel, 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. The present report provides information about the design, administration, and scoring of the test.

Purpose Statement

The purpose, or intended use, of the Knowledge for Teaching Elementary Fractions (K-TEF) test is to produce ability estimates that can be used to serve as the dependent variable in models estimating the effect of the intervention on teacher MKT, as well as to investigate MKT as a potential mediator of the effect of the program on 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. Because the K-TEF administered in spring 2017 as a postintervention measure of teacher knowledge and abilities is very similar to the K-TEF administered in fall 2016 as a preintervention measure, much of the text in the two corresponding reports is similar.

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 contained an initial 19 items, prompting up to 27 individual responses from the test taker. Twenty-one of the 27 responses used a selected-response format (including 10 yes/no responses), and the remaining 6 a constructed-response (fill-in-the-blank) format. After evaluation of the psychometric properties of the test and items, 2 items were dropped; 17 items—yielding 25 responses—contributed to the final scale.

Sample and Setting

The test was administered to a sample of 241 elementary educators in six U.S. states in spring 2017. One examinee responded to less than 75% of the possible responses and was dropped from analysis, leaving an analytic sample of 240 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.

Dimensionality

To investigate the dimensionality of the test data, we performed parallel analysis using the final-scale (17-item) format. Results of these analyses suggested the K-TEF test measures a single construct.

Item Diagnostics and Scoring

Item diagnostics and calibration accounting resulted in collapse of the 27 individual responses (or nonresponses) into 19 independent items. After two of the constructed-response items were removed because of poor psychometric outcomes, the remaining 25 were included in the final 17-item scale.

Initial screening of the items used an approach based on classical test theory (CTT). When the CTT-based p-value was used to estimate item difficulty, the mean difficulty estimate for the 17 items in the final scale was .58, the minimum value was .82, and the maximum value was .14, suggesting a broad range of difficulty among items on the test. The mean item-rest correlation coefficient was .38, the minimum value was .28, and the maximum value was .47, suggesting that the items in the final scale had adequate discriminative power.

Item-Response Theory Data Modeling

Because the test form included 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 0.79 to 1.67 (M = 1.16, SD = 0.26).

Maximum-likelihood estimator and *expected a posteriori* (EAP) 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 EAP estimator.

Reliability and Test Information

By means of a CTT approach, coefficient α and standard error of measurement (SEM) were calculated to be .77 and 2.21, respectively. In addition, test information and conditional standard error of measurement (CSEM) were generated through an IRT-based approach. Test information quantifies the degree of accuracy for every ability-score (θ) level. The higher it is, the more accurately the ability is estimated. Test information and CSEM has a reciprocal relationship; when one goes high, the other goes low. High test information implies low CSEM, and low test information implies high CSEM. The highest test information and the lowest CSEM occurred when the person ability (θ) was approximately –0.40. The person-ability estimate was associated with higher test information and lower CSEM for the person ability estimates between –1.20 and 0.00 on the θ scale and was associated with lower test information and higher CSEM for the person-ability estimates greater than 2.00 on the θ scale.

Distribution of Educator Ability Scores

Using the EAP estimator, we found that the distribution of the freely estimated educator ability (θ) scores for the educators in the present sample did not appear to differ from a normal distribution. By the EAP method, the θ estimates for the educators in the sample ranged from –1.95 to 2.28 (M = 0.00,

SD = 0.89). The skewness and the kurtosis statistics for the sample distribution were 0.17 and –0.68, respectively.

Test-Score Equating

The K-TEF administered to teachers in fall 2016 (Schoen, Yang, Liu, & Paek, 2018) served as a preintervention measure of mathematical knowledge for teaching for educators who were participating in a randomized trial. The K-TEF administered to teachers in spring 2017—the focus of our report—served as a postintervention measure. The items and responses contributing to the final scale for the two tests have 16 items in common and one item that was used only on the postintervention test. Consequently, the pre- and postintervention test scores are not directly comparable. The preintervention covariate need not be on the same scale to permit inferences about the effect of an intervention on the outcome of interest in a randomized-controlled trial, but being able to compare postintervention and preintervention scores directly is sometimes advantageous. To that end, we conducted a brief equating study using a fixed-item-parameter approach. The common items on the pre- and postintervention tests were used as anchor items; they provided the link between the two tests. The anchor-item parameter estimates created from the preintervention sample were fixed in the postintervention data calibration, so that the pre- and postintervention θ estimates (and SEMs) would be on the same scale.

When the EAP method for θ estimation on the equated scale was used, the postintervention person-ability scores ranged from –1.84 to 2.32. The mean and standard deviation of the EAP estimates were .09 and .86, respectively. The skewness and the kurtosis of the distribution of person-ability estimates were .16 and –.63, respectively. Person-ability estimates around the value of .00 on the equated scale were associated with the highest test information and the lowest CSEM. In addition, the person-ability estimates were related to lower CSEM (i.e., more accurate estimation of person ability) when they ranged between –1.00 and 1.00. Person-ability estimates were related to higher CSEM (i.e., less accurate estimation of person ability) when they were larger than 2.00 or less than –2.40.

Predictive Validity

We calculated the Pearson correlations of teacher preintervention and postintervention from a sample of 228 teachers who completed both tests. The correlation of the equated θ estimates using the EAP method was .67.

Discussion and Conclusions

In summary, we found evidence that the Knowledge for Teaching Elementary Fractions test measures a single construct. Reliability, test-information, and item-discrimination estimates appear to fit the intended purpose of the test, although further validation will be necessary to reveal whether the test is well suited for its intended use. Evaluation of the structural validity of the resulting 17-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 (K-TEF) test. In terms of the mathematical knowledge for teaching (MKT) theoretical framework, the items on this test that comprise the final score were designed to measure common content knowledge and specialized content knowledge (Ball, Thames, & Phelps, 2008) on the topic of fractions.

The items on the test were designed to measure fractions-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 were not designed to create subscales. Rather, the test was 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 (Saderholm, Ronau, Brown, & Collins, 2010), Learning Mathematics for Teaching project (Hill, Schilling, & Ball, 2004; LMT, 2004), Numeracy Development Projects (Ward & Thomas, 2015), the Knowledge for Teaching Elementary Mathematics project (Schoen, Bray, Wolfe, Tazaz, & Nielsen, 2017), the Mills College Lesson Study Group (https://lessonresearch.net), 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) and as the pretest in the present study. The previous version of the K-TEF test detected a significant difference between teachers in a treatment group and those in a control group (Lewis & Perry, 2017).

The version of the test used for the present sample was almost identical to the preintervention K-TEF, which was administered in fall 2016. One four-part item set (i.e., testlet) on the K-TEF form used in fall 2016 was removed before the K-TEF form used in spring 2017 was administered. It corresponds to items 3a, 3b, 3c, and 3d, or item 2*, in the fall 2016 K-TEF (Schoen, Yang, Liu, & Paek, 2018). This item was dropped, because of the content analysis suggested poor alignment with the intervention, which emphasized linear representations of fractions. One item that was not used on the 2016 K-TEF form was added to the spring 2017 K-TEF form. It was drawn from the Knowledge for Teaching Elementary Mathematics (Schoen, 2018) item bank and asks the examinees to solve a word problem involving a length-measurement scenario. The given numbers given in the word problem are whole-number lengths, and the correct answer is a fractional length.

1.1. Description of the Sample

The present report focuses on the version of the K-TEF test that was administered to a group of 241 educators from six states in the U.S. in spring 2017. Characteristics of the individuals in the sample are provided in Table 1.1. Approximately 83% of the sample were regular classroom teachers, the majority of whom were teaching third, fourth, or fifth grade. The average reported number of years of teaching experience among teachers in the sample was 13.3

Table 1.1. Characteristics of Teachers in the Spring 2017 Field-Test Sample (N = 241)

Characteristic	Total (proportion)
Primary teaching role	
Regular classroom[a]	200 (.830)
Varying exceptionalities[b]	15 (.062)
English language learners	3 (.012)
Other[c]	22 (.091)
Departmentalization	
Teaches all subjects	159 (.660)
Teaches only mathematics	74 (.307)
Does not teach mathematics	7 (.029)
Grade level primarily taught	
Kindergarten	0 (.000)
Grade 1	2 (.008)
Grade 2	12 (.050)
Grade 3	100 (.415)
Grade 4	77 (.320)
Grade 5	41 (.170)
Grade 6	8 (.033)
Grade 7	0 (.000)
Grade 8	0 (.000)
Highest degree earned	
No degree	0 (.000)
Associate's degree	1 (.004)
Bachelor's degree	121 (.502)
Master's degree	95 (.394)
Specialist degree	23 (.095)
Areas of certification	
Elementary Education	214 (.888)
PreK/Primary Education	36 (.149)
Middle Grades Mathematics	20 (.083)
Secondary Mathematics	5 (.021)
ESOL/Bilingual/Dual-language	102 (.423)
Varying Exceptionalities[b]	59 (.245)
State[d]	
Florida	161 (.672)
Illinois	29 (.120)
California	28 (.116)
Colorado	7 (.029)
Indiana	3 (.012)
New York	12 (.050)
Years of teaching experience	13.3 ± 7.6

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 only responded to confirmation of location (state) ,leaving 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 K-TEF test. All the items on this test were borrowed or adapted from other sources. An account of the source of each item is provided in Appendix A.

Table 1.2. Test Blueprint for the Spring 2017 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*
Maria needs to swim miles	3	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	
Steve- fiction is more than Andrew fiction. Correct?	8a	6	4*
Why/why not is Steve not correct?	8b		
Highway under construction	9	7	5*
	10a	8	6*
	10b	8	6*
	10c	8	6*
Given yards rope, with per rope, how many ropes?	11	9	7*
Student representations of	12	10	8*
Jim's proportion of program sessions taught	13	11	9*
Word problem for); pizza	14a	12	10*
Word problem for ; corn field	14b	12	10*
Word problem for ; sugar	14c	12	10*
Word problem for ; licorice	14d	12	10*
Divide equally among students	15	13	11*
	16	14	12*
Models to represent	17	15	13*
Connections- measurement and fractions	18		
Fractional part of square in triangle A	19	16	14*
Paper frog moving along a line	20	17	15*
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 za	24a	18	16*
Word problem	24b	18	16*
Word problem	24c	18	16*
Word problem	24d	18	16*
Comparing	25	19	17*

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 exclusion of qualitative items and formation of 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 K-TEF test includes 25 items, which prompt 35 responses from examinees, because some items (i.e., item 1, 8, 10, 14, and 24) are testlets that require multiple responses. (See Appendixes A and B for specifics.) The 35 responses can therefore be split into two groups, of which the first consists of 27 responses that can be scored as correct or incorrect. These correspond to both selected-response and 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 K-TEF 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 27 fraction-focused responses in the recoded test were scored dichotomously as correct or incorrect in accordance with the answer keys. Because recoded items 10, 14, 24 required multiple responses, we scored those items polytomously by summing the scores of their responses within each testlet. The recoding was intended 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 items 1 and 5[1] (by the recoded item numbering scheme) from the recoded test. The final version of the test therefore consisted of 17 items. We placed an asterisk after the item numbers on the final test item numbers to avoid confusion with the item numbers on the original test form or 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.

[1] The reasons for removing items 1 and 5 are provided in section 3.3.

3. Data and Scoring

3.1. Data Entry and Verification Procedures

The K-TEF test was administered as an on-line survey using Qualtrics software. Response data were exported from Qualtrics to a flat file and manipulated by means of SPSS and Excel software.

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 produced missing item-level responses and multiple submissions for individual participants. When individual participants submitted multiple responses for a given item, the one with the latest date was taken to be the final response.

3.2. Missing Data

Because examinees were allowed to skip items in the web-based questionnaire, some examinees provided incomplete responses. Table 1 shows the missing response(s) in the sample. The original sample size was 241. After review of patterns in the missing data, a decision was made to exclude cases with response rates lower than 75%.

3.2 Missing Data

Because examinees were allowed to skip items in the web-based questionnaire, some examinees provided incomplete responses. Table 1 shows the missing response(s) in the sample. The original sample size was 241. After review of patterns in the missing data, a decision was made to exclude cases with response rates lower than 75%. One examinee completed only the first six responses and did not respond to any of the remaining items (i.e., 77.78% missing data); this examinee's response data were excluded from the following data analyses. The remaining nonresponses were scored as incorrect responses. After exclusion of that one examinee, the final sample size was 240.

Table 3.1. Missing Response Frequency in the Sample

No. of missing response(s)	Frequency	% of sample	Cumulative %
0	194	80.5	80.5
1	27	11.2	91.7
2	12	5.0	96.7
3	3	1.2	97.9
4	1	0.4	98.3
5	2	0.8	99.2
6	1	0.4	99.6
21	1[†]	0.4	100.0
Total	241	100.0	

Note. # of Missing response(s) = the number of missing response(s) for a given examinee in the sample; frequency = the number of examinees with a given number of missing response(s); % of sample = the percentage of examinees who had given numbers of missing response(s); cumulative % = cumulative percentage of examinees who had given numbers of missing response(s).
[†]This examinee's data were excluded from the analysis.

3.3. Item Scoring

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 27 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 A.

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.

Some items prompted multiple responses from the same item stem. For example, item 14 of the original test requires four responses from teachers, and teachers' scores on item 14 are represented by a polytomous variable defined as the sum of four dichotomous variables, corresponding to the four responses (see Table 1.2). We decided to score these three items polytomously as a means to address concerns about local dependence when testlets were used and item-response theory was used to estimate teachers' latent abilities.

After excluding qualitative items and forming the polytomous items, we further excluded items 1 and 5 for the following reasons. First, when conducting the parallel analysis, we used the *cor.smooth* (Revelle, 2017) function of the *psych* (Revelle, 2017) package in R 3.4.0 (R Core Team, 2012) and determined that items 1 and 5 were causing problems in the estimation of polychoric correlations. Second, the results of the the CTT-based analyses, which were performed after the dimensionality test (see Section 4), indicated the corrected item-total (i.e., item-rest) correlation coefficients of these two items were low. The item-rest correlation coefficient for item 1 was .18; that for item 5 was .21.

After items 1 and 5 were excluded, the final-scale test had 17 items (see Table 1.2). The remainder of the present report focuses on results from analysis of the final-scale test.

4. Dimensionality Analysis

Parallel analysis is a procedure for examining the number of constructs in the data, 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). We conducted parallel analysis to examine the dimensionality of the test using the final-scale items. The *psych* (Revelle, 2017) program in R 3.4.3 (R Core Team, 2017) was used to perform the analysis.

Figure 4.1 shows the results of the parallel analysis. The vertical axis in the figure represents the eigenvalues of the principal components, and the horizontal axis represents the number of components. The blue line is for the principal components from the actual data, and the dotted red line is for the principal components from the simulated (or resampled) data. The number of components from the actual data above the dotted red line indicates the number of dimensions in the data. The decision was made in consideration of the confidence intervals (which are shown as the vertical error bars in the figure) for the resampled data. The results suggested that the test was measuring a single construct.

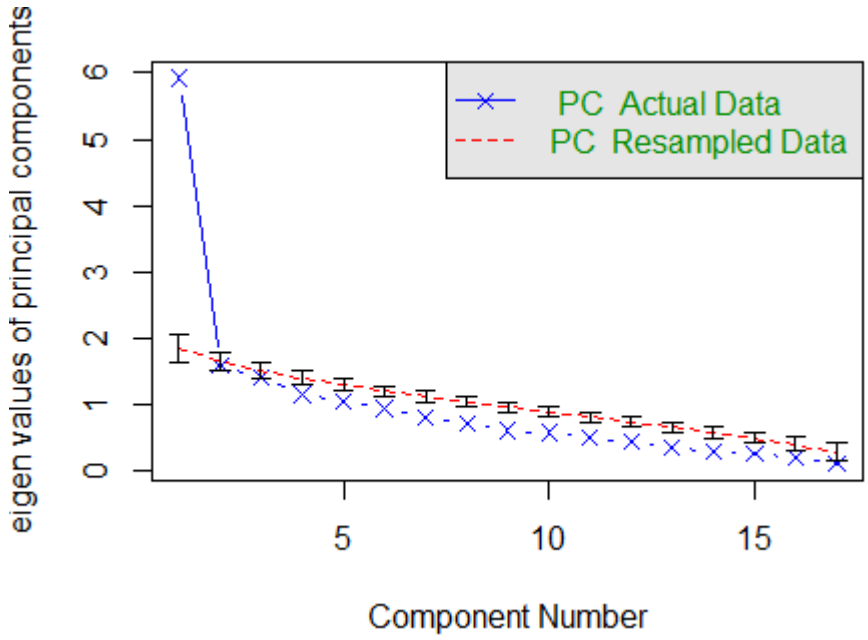

Figure 4.1. Parallel analysis scree plot. PC, principal components.

5. Classical Testing Theory (CTT) Analyses

5.1. Distribution of the Observed Test Score

After finding that the parallel analysis suggested that the test was unidimensional, we conducted the CTT analyses using SPSS 22.0 (IBM Corp., 2013). Figure 2 displays the distribution of the observed total raw scores in the final-scale format. The mean of the observed test scores was 15.25, and the standard deviation was 4.61. The median was 15.00, the skewness was –0.05, and the kurtosis was –0.83. Note that although the final-scale format had only 17 items, the observed test scores ranged from 5.00 to 25.00, because items 6*, 10*, and 16* were coded into polytomous items.

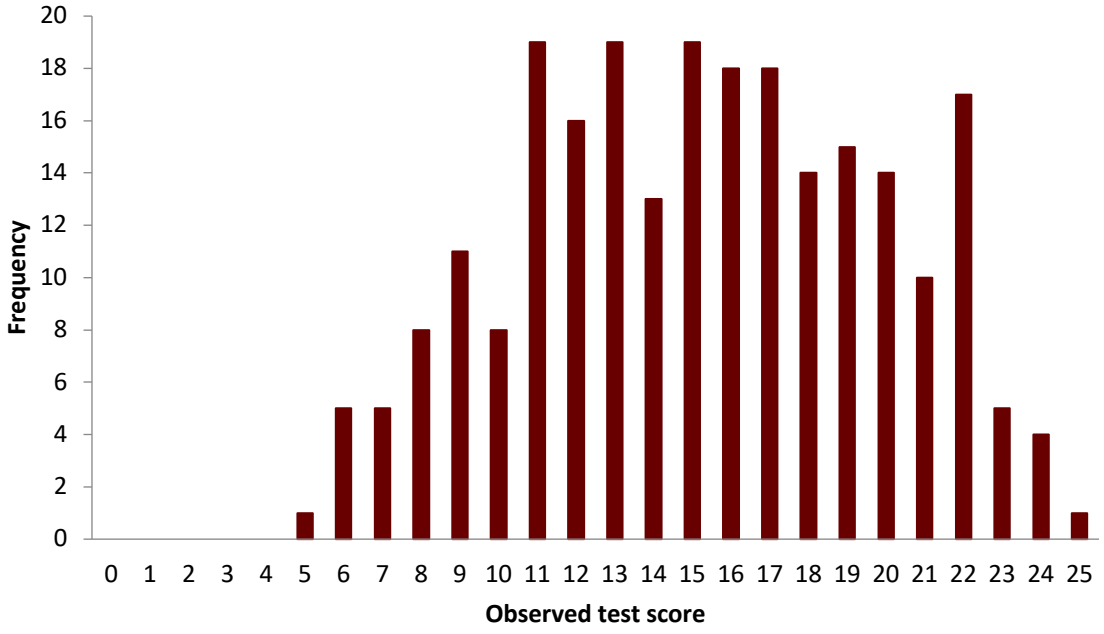

Figure 5.1. Distribution of the observed test scores in the final-scale format.

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. Tables 5.2 and 5.3 show the distribution of item difficulty and item discrimination for the 17 items used in the final scale. Figures 5.2 and 5.3 display the item-difficulty and item-discrimination estimates, respectively, for each sequential item on the final-test scale.

The item-difficulty statistics ranged from a maximum difficulty estimate of .14 (item 8*) to a minimum difficulty estimate of .82 (item 6*). The mean of the item-difficulty estimates was .58 with a standard deviation of .19. The skewness statistics of the item-difficulty estimates in the test was –.61, and the kurtosis statistics was .07.

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 ranged from a minimum of .28 (item 8*) to a maximum of .47 (item 12*). The individual item-discrimination estimates were greater than .20 for every item in the final-scale test. The mean of the item-discrimination estimates was .38, with a standard deviation of .05. The skewness statistic was –.03, and the kurtosis statistic was –.67.

Table 5.1. Item Difficulty and Discrimination Estimates Based on CTT Analyses

Final-scale item #	*M*	*SD*	*p*	Item-rest *r*
1*	0.55	0.50	.55	.39
2*	0.78	0.42	.78	.35
3*	0.77	0.42	.77	.42
4*	0.75	0.44	.75	.30
5*	0.39	0.49	.39	.40
6*	2.47	0.80	.82	.41
7*	0.77	0.42	.77	.33
8*	0.14	0.35	.14	.28
9*	0.50	0.50	.50	.46
10*	2.52	0.96	.63	.36
11*	0.55	0.50	.55	.43
12*	0.48	0.50	.48	.47
13*	0.51	0.50	.51	.33
14*	0.78	0.41	.78	.39
15*	0.43	0.50	.43	.36
16*	2.52	1.13	.63	.43
17*	0.36	0.48	.36	.33

Note. Final-scale item # = coding testlets polytomously and removing two problematic items (we differentiated recoded item index and final-scale item index by placing a * after the final-scale item number); *p* = item difficulty; Item-rest *r* = item-rest correlation coefficient (i.e., corrected item-total correlation coefficient), which is the Pearson correlation between the item score and the test score that excludes the item score.

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	0
.80–.89	1
.70–.79	5
.60–.69	2
.50–.59	4
.40–.49	2
.30–.39	2
.20–.29	0
.10–.19	1
.00–.09	0
Mean	.58
Standard Deviation	.19
Maximum	.82
Minimum	.14

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	7
.20–.39	10
.00–.19	0
Mean	.38
Standard Deviation	.05
Minimum	.28
Maximum	.47

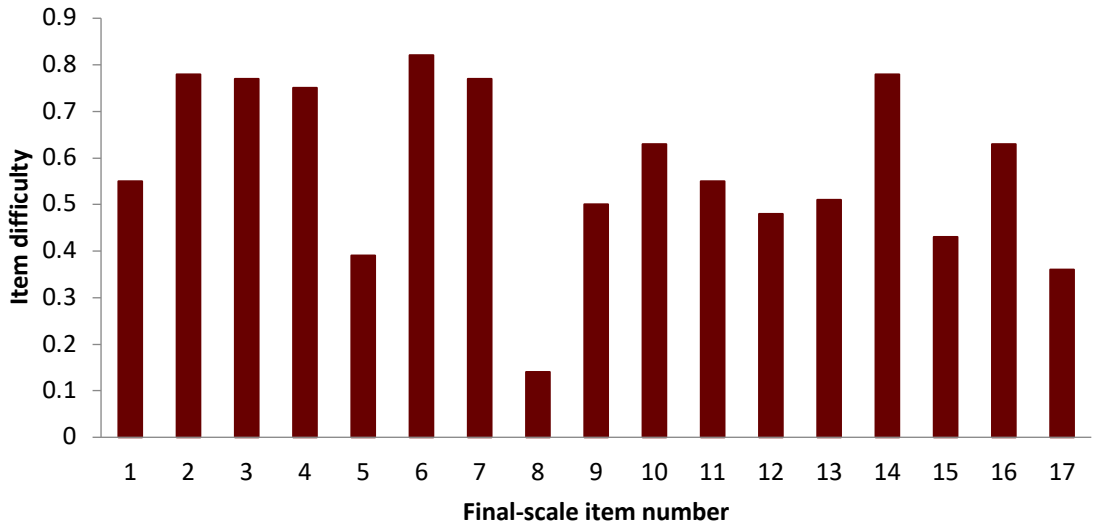

Figure 5.2. Item difficulty estimate of each final-scale item.

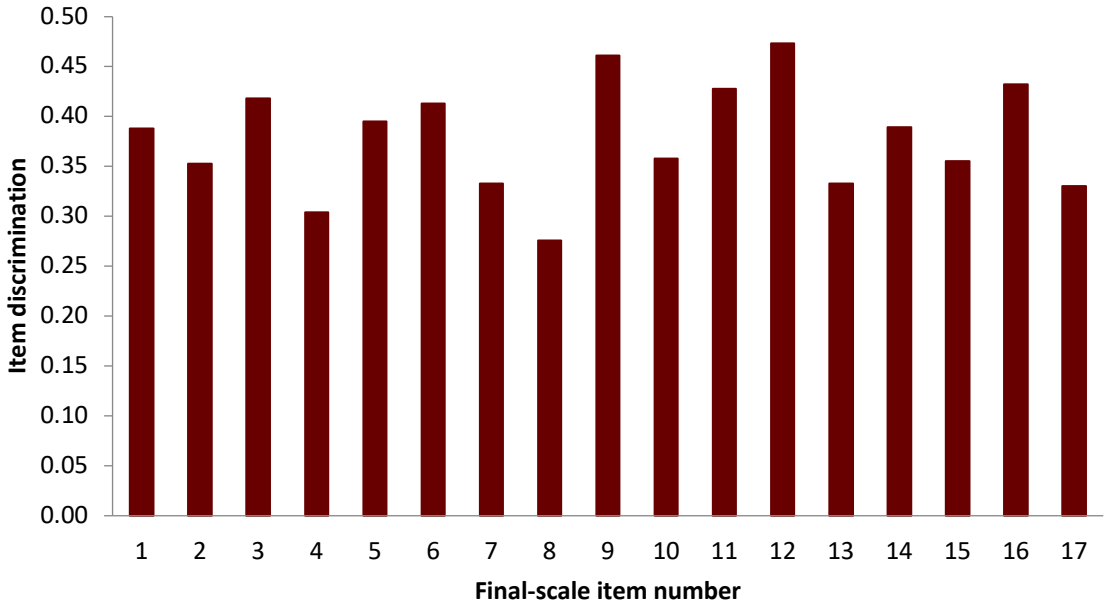

Figure 5.3. Item discrimination estimate of each final-scale item.

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 .77. SPSS output indicated that the scale variance was 21.25. We subsequently calculated the standard error of measurement (SEM) of the test. On the basis of Equation 2,

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

where σ^2 is the test variance, and ρ_{XX} is the reliability of the test, SEM was calculated to be 2.21.

6. Item Response Theory (IRT) 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*, 2*, 3*, 4*, 5*, 7*, 8*, 9*, 11*, 12*, 13*, 14*, 15*and 17*), a two-parameter (2PL) model was used. For the polytomous items (6*, 10*, 16*), 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 5433.28. 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* $\in \{0, 1, 2, \dots, k\}$, 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 17 items ranged from 0.79 (item 10*) to 1.67 (item 14*). Twelve of the 17 items in the final scale had discrimination estimates greater than 1.0. The mean of the item discrimination parameter estimates for all 17 items was 1.16, with a standard deviation of 0.26. The skewness statistic was 0.55, and the kurtosis statistic was –0.55. For the 14 items using 2PL models, the item-difficulty estimates ranged from a minimum of –1.42 (item 6*) to a maximum of 2.15 (item 7*). The mean of the item-difficulty estimates was –0.21 with a standard deviation of 0.99. The skewness statistic was 0.81 and the kurtosis statistic was 1.01.

Table 6.1. Parameter Estimates and Standard Errors for Final-Scale Items Modeled with Two-Paramater Logistic Methods

Item #	*a (SE)*	*b (SE)*
1*	1.09 (0.24)	–0.20 (0.17)
2*	1.33 (0.28)	–1.26 (0.23)
3*	1.47 (0.31)	–1.13 (0.22)
4*	1.06 (0.29)	–1.25 (0.31)
5*	1.26 (0.26)	0.45 (0.16)
7*	1.03 (0.27)	–1.42 (0.34)
8*	1.01 (0.27)	2.15 (0.46)
9*	1.57 (0.30)	0.00 (0.13)
11*	1.29 (0.26)	–0.24 (0.15)
12*	1.32 (0.28)	0.09 (0.15)
13*	0.85 (0.22)	–0.05 (0.20)
14*	1.67 (0.43)	–1.14 (0.21)
15*	0.91 (0.21)	0.38 (0.20)
17*	0.97 (0.23)	0.71 (0.21)

Note. Final-Scale Item # = coding testlets polytomously and removing two problematic items (we differentiated recoded item index and final-scale item index by placing a * after the final-scale item number); *a* = item discrimination index; *b* = item difficulty index; *SE* = standard error.

Table 6.2. Parameter Estimates and Standard Errors for Final-Scale Items Modeled with the Graded Response Model

Item #	*a (SE)*	b_1 *(SE)*	b_2 *(SE)*	b_3 *(SE)*	b_4 *(SE)*
6*	1.13 (0.24)	–3.58 (0.72)	–2.00 (0.40)	–0.60 (0.19)	
10*	0.79 (0.18)	–5.24 (1.30)	–2.29 (0.52)	–0.47 (0.23)	2.76 (0.60)
16*	0.94 (0.21)	–4.12 (1.01)	–1.72 (0.42)	–0.15 (0.18)	1.43 (0.29)

Note. Final-Scale Item # = coding testlets polytomously and removing two problematic items (we differentiated recoded item index and final-scale item index by assigning a * after the final-scale item number); *a* = item discrimination index; b_{jk} $(j = 1,2,\ldots,17, k = 0,1,2,3,4)$ = category threshold; *SE* = standard error.

6.3. Test Information and Estimated Person Ability

Figure 6.1 displays the test information curve and conditional standard error of measurement (CSEM) for the K-TEF 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 (i.e., θ) estimate around the value of –0.40 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 –1.20 and 0.00; the curve also suggested that person-ability estimates were related to higher CSEM (i.e., less accurate estimation of person ability) when the person ability was higher (e.g., around 2.00).

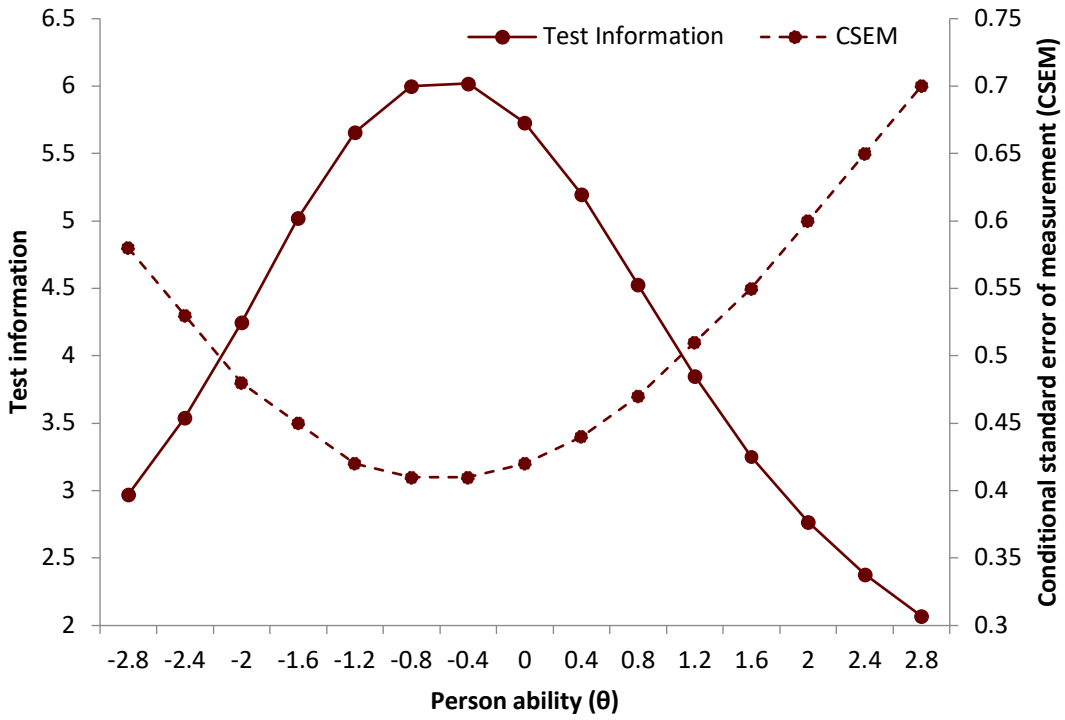

Figure 6.1. Test information curve and CSEM for the final-scale items.

Figure 6.2 presents the person-ability estimates produced by the maximum likelihood estimation (MLE) method. For individuals who got perfect or zero scores, the ability estimates based on the MLE method are not available. As shown in Figure 6.2, a spike appeared at higher end of the horizontal axis, because one examinee had a perfect score for the test. When person ability was estimated by MLE, the minimum- and the maximum-likelihood scores were set at –7 and 7, respectively, in the flexMIRT software.

We also used the expected *a posteriori* (EAP) method to estimate person ability. Figure 6.3 presents the resulting distribution of person-ability estimates. The person-ability estimates ranged from –1.95 to 2.28. The mean and standard deviation of the EAP estimates were 0.00 and 0.89, respectively. The skewness and kurtosis of the person-ability distributions were 0.17 and –0.68, respectively.

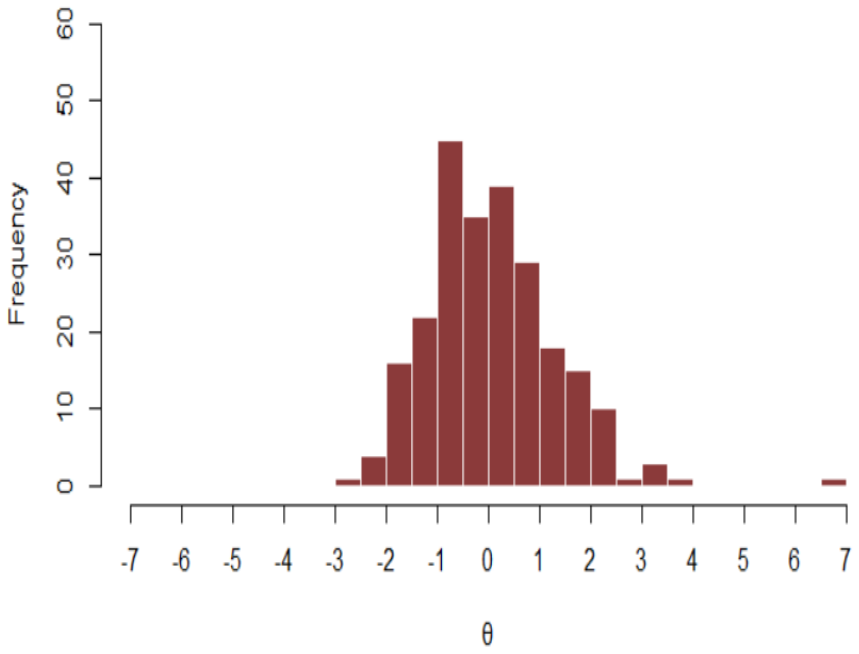

Figure 6.2. Sample distribution of person abilities (θ) estimated by maximum-likelihood estimation.

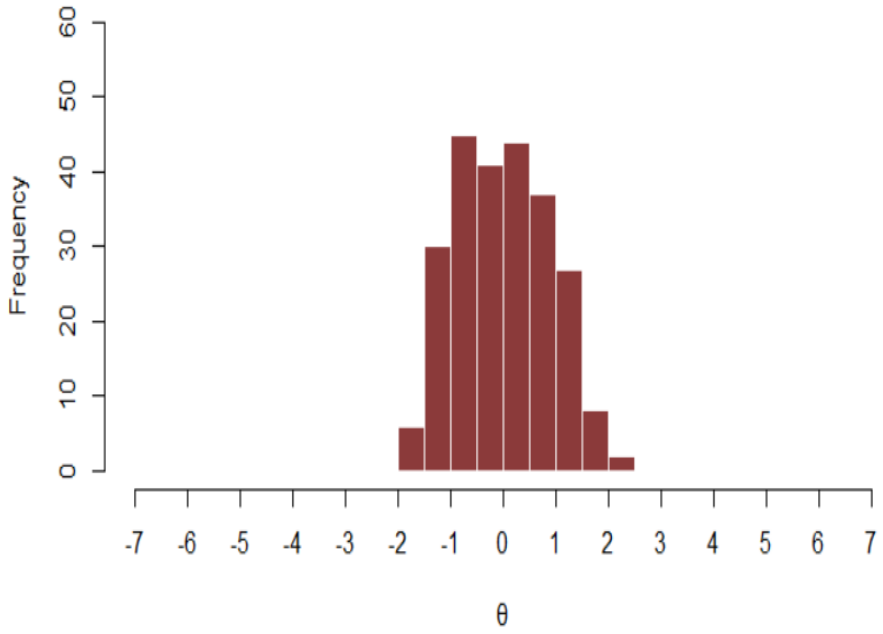

Figure 6.3. Sample distribution of person abilities (θ) estimated by expected a priori *method.*

6.4. Equating Posttest and Pretest Scores

The K-TEF was administered to teachers in spring 2017 as a post-intervention measure of teachers' ability in a randomized trial of an educational intervention involving lesson study and fractions resource kits. The K-TEF administered to teachers in fall 2016 (Schoen, Yang, Liu, & Paek, 2018) served as a preintervention measure of many of those same educators' abilities. The model-parameter estimates and the θ scores reported here thus far have not been on the same scale as those for the preintervention test, so the pre- and postintervention test scores are not directly comparable. The preintervention test covariate need not be on the same scale to permit inferences about the effect of an intervention on the outcome of interest in a randomized controlled trial, but being able to compare post- and preintervention test scores directly is sometimes advantageous. We therefore conducted a brief equating study.

6.4.1. Equating Process

Equating is a statistical process for adjusting scores of different test forms, in this case, pretest and posttest, so we can directly compare scores on the pretest and posttest test forms (Kolen and Brennan, 2004). There are several different methods for equating of observed scores, such as traditional equating methods (i.e. mean equating, linear equating, equipercentile equating) as well as IRT equating methods.

To equate teacher pre- and postintervention K-TEF scores, we used a fixed-item-parameter approach (Kang and Petersen, 2009) with separate calibrations. The common items on the two tests were used as anchor items and provided the link between them. The anchor-item parameter estimates created from the preintervention sample were fixed in the postintervention data calibration, so the pre- and postintervention θ estimates (and SEMs) were on the same scale.

The final scale for the preintervention test included 18 items, and that for the postintervention test included 17 items. The two tests had 16 items in common, and the post-intervention tests included, as well, one unique item. We therefore fixed the item parameters of the 16 common items in the postintervention test at the same values as the corresponding item-parameter estimates used in the final-scale scoring of the preintervention data gathered in fall 2016. Meanwhile, the item parameters for the one unique item on the postintervention test, as well as the population mean and variance, were freely estimated from the data gathered in spring 2017.

6.4.2. Item Difficulty and Discrimination

Tables 6.3 and 6.4 present item-parameter values of equated teacher postintervention test final-scale items. The equating assumes the anchor items are invariant, so the 16 anchor-item parameter values were the same on pre- and postintervention tests. The item parameters for the one unique item (item 2* in the postintervention-test final scale) were freely estimated. More information about the K-TEF test used in fall 2016 and the corresponding item-parameter estimation is available in Schoen, Yang, Liu, & Paek (2018).

Table 6.3. Item-Parameter Values of Equated Postintervention-Test Final-Scale Items Modeled from Two-Parameter Logistic Methods

Pretest final-scale #	*Posttest final-scale #*	*a*	*SE*	*b*	*SE*
1*	1*	1.20	–	0.07	–
	2*	1.36	0.27	–1.13	0.21
3*	3*	1.13	–	–0.63	–
5*	4*	1.26	–	–1.08	–
6*	5*	1.77	–	0.26	–
8*	7*	1.32	–	–1.38	–
9*	8*	1.44	–	1.85	–
10*	9*	1.65	–	0.16	–
12*	11*	0.91	–	–0.46	–
13*	12*	0.89	–	0.19	–
14*	13*	0.77	–	–0.05	–
15*	14*	1.17	–	–1.15	–
16*	15*	1.32	–	0.20	–
18*	17*	0.78	–	0.87	–

Note. Final-Scale # = the newly generated item number after formation of polytomously scored item and removal of problematic item (asterisks follow item numbers used in the final scale); *a* = item discrimination index; *b* = item difficulty index; *SE* = standard error. Item 2* in the posttest final scale is a unique item in the posttest, and all the other 16 items in the posttest final scale are the common items between pretest and posttest.

Table 6.4. Item Parameter Values of Equated Posttest Final-Scale Items Modeled Using the Graded Response Model

Posttest final-scale #	*Posttest final-scale #*	*a*	*SE*	*b1*	*SE*	*b2*	*SE*	*b3*	*SE*	*b4*	*SE*
7*	6*	0.82	–	–4.96	–	–2.16	–	–0.35	–		–
11*	10*	0.85	–	–3.73	–	–2.56	–	–0.27	–	2.89	–
17*	16*	1.02	–	–2.62	–	–1.55	–	–0.03	–	1.43	–

Note. Final-Scale Item # = the newly generated item number after forming polytomously scored item and removing problematic item (asterisks follow 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. Item 2* in the posttest final scale is a unique item in the postintervention test, and all the other 16 items in the postintervention test final scale are the same as those in the preintervention test.

6.4.3. Test Information and Estimated Person Ability

Figure 6.4 displays the test information curve and the conditional standard error of measurement (CSEM) for the equated-scale test. According to the relationship between test information and CSEM, a person ability (θ) estimate around the value of .00 was associated with the highest test information and the lowest CSEM. In addition, the person-ability estimates were related to lower CSEM (i.e., more accurate estimation of person ability) when they ranged between –1.00 and 1.00; the curve also suggested that person ability-estimates were related to higher CSEM (i.e., less accurate estimation of person ability) when they were higher than 2.00 or below –2.40.

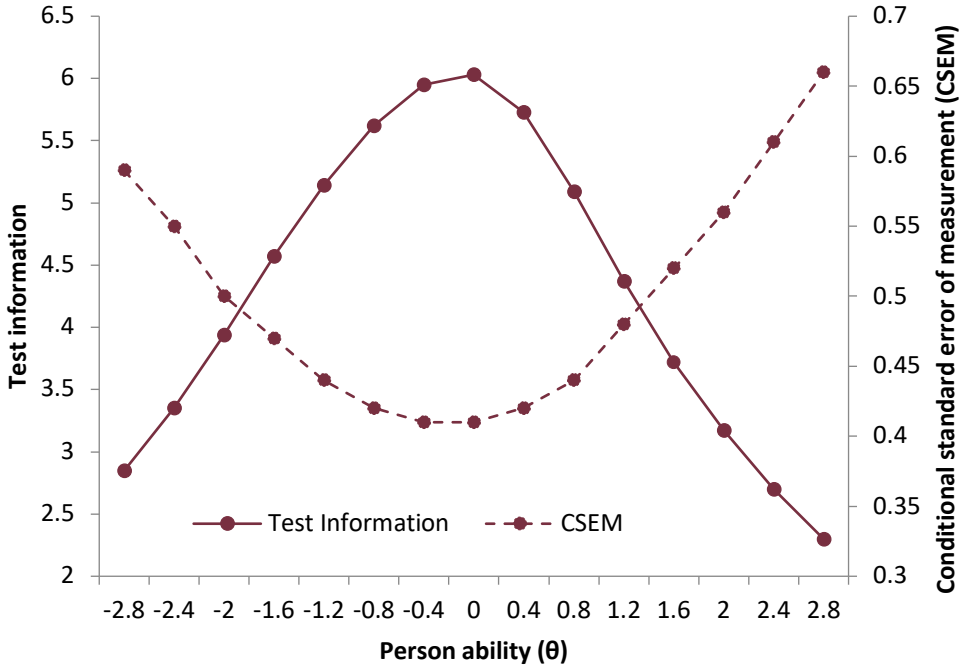

Figure 6.4. Test information curve and conditional standard error of measurement for the equated posttest final scale.

Figure 6.5 presents the equated person-ability estimates for the postintervention test scores in the spring 2017 sample according to MLE. A spike appeared at the higher end of the horizontal axis, because one examinee had a perfect score for the test. When person ability was estimated by MLE, the minimum and the maximum likelihood scores were set at –7 and 7, respectively, in the flexMIRT software.

We also used the EAP method to estimate person ability for the scores in the spring 2017 sample. Figure 6.6 presents the distribution of equated person ability on the equated pre- and postintervention scale using the EAP method. The person-ability scores ranged from –1.84 to 2.32. The mean and standard deviation of the EAP estimates were .09 and .86, respectively. The skewness and the kurtosis were 0.16 and –0.63, respectively.

6.4.4. Predictive Validity

From a sample of 228 teachers who completed both the pre- and post-intervention tests, the Pearson correlations of the two tests were calculated by three different methods. The correlation based on common-item observed test scores was .67. According to equated θ estimates from the ML method, the correlation was .62, and that from the EAP method was .67.

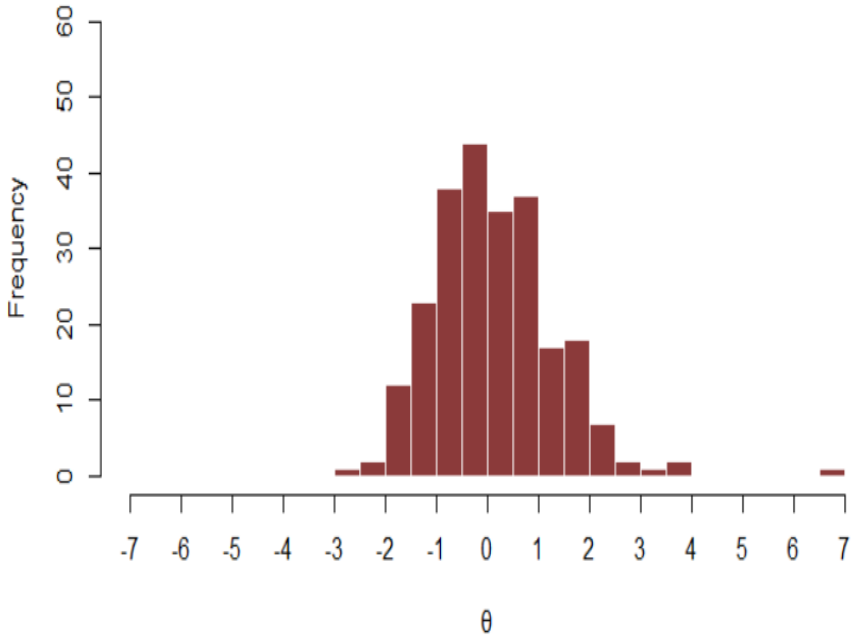

Figure 6.5. Equated Person abilities (θ) estimated by maximum likelihood estimation.

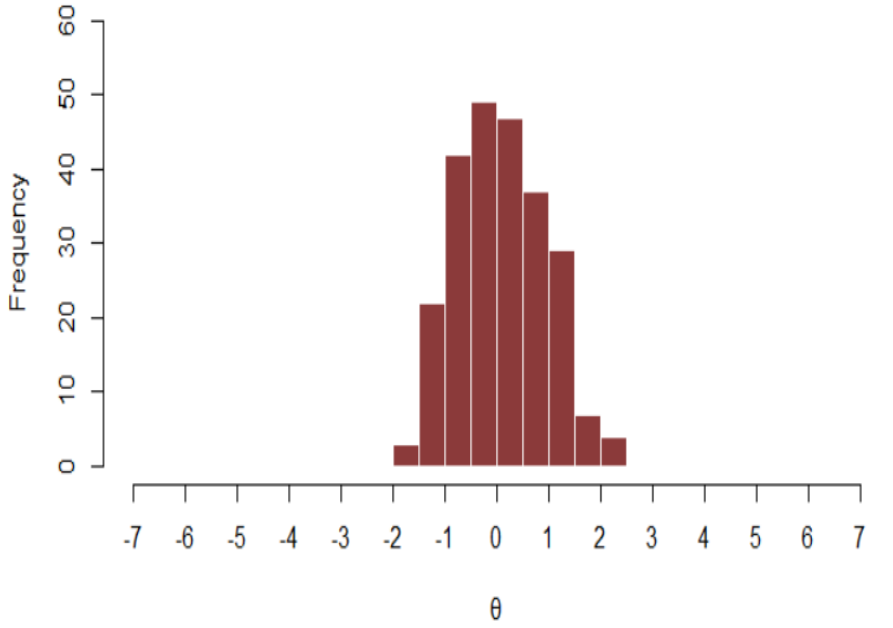

Figure 6.6. Equated Person abilities (θ) estimated by the estimated a priori *method.*

7. Discussion

Here we report findings from a field test of the Knowledge for Teaching Elementary Fractions (K-TEF) test during spring 2017. This psychometric report contributes to the validation of the test in several important ways. The discussion below is 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 sources. Each source was subjected to review by content experts and/or peer review before publication in refereed journals. 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. Because the test was found to measure a unidimensional construct, subcategories were not be necessary, but they might provide additional description and support for the interpretation of scores. For example, the items could be sorted according to referent unit, partitioning and iterating, and relative magnitude of fractions. They could also be divided according to content and pedagogical content knowledge or by domains within more specific theoretical frameworks for MKT (Ball, Thames, & Phelps, 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. Additional research in this area could yield insight into the facets of knowledge that are affected by the intervention or associated with student achievement.

7.2. Structural Validity

7.2.1. Unidimensionality

Parallel analysis indicated the test measured a single construct, which is consistent with previous field test of a nearly identical test (Schoen, Yang, Liu, & Paek, 2018). Within the theoretical framework of mathematical knowledge for teaching (Ball, Thames, & Phelps, 2008), we assert this construct to be identified as mathematical content knowledge. We do not assert that the K-TEF measures pedagogical content knowledge in mathematics, per se, or pedagogical knowledge as defined by Shulman (1986).

7.2.2. Level of Difficulty for the Intended Population

The difficulty of the test was reasonably well aligned with the ability level of the educators in the sample, but the peak test information occurred at an ability level of –.4 when the postintervention scores were freely estimated from the test data. A nearly identical test had been administered to the same subjects approximately eight months earlier. Approximately three-quarters of the subjects participated in an educational intervention between the two administrations of the test. On the basis of the preintervention test data and the EAP estimator, the sample mean ability estimate for the preintervention sample was 0.00. When the postintervention data and the EAP estimator were used to freely estimate item parameters, variance, and population means, the sample mean ability estimate for the postintervention sample was 0.06. The ability estimates for the participants in both samples were approximately normally distributed. After a fixed item parameter-equating method based on item-

response theory was applied, the sample mean ability estimate for the postintervention sample was 0.09, indicating that the overall sample had higher ability at after the intervention than before it.

The equating procedure was based on an assumption of no item drift. The item-parameter estimates based on the pre- and postintervention sample data were not identical, but they were reasonably close to one another. Because an intervention came between the two, and a very similar intervention has already been found to have a positive effect on educators' MKT (Lewis & Perry, 2017), many of the examinees' abilities could reasonably be expected to have increased between the two tests. These changes in ability may be the cause of the shift in the sample mean-ability estimates and the shift in the range of ability estimates with high test information in the freely estimated postintervention scale from that of the preintervention scale.

Several plausible explanations are possible for the difference in alignment of test difficulty with person ability. The intervention may have caused an increase in the performance of some of the examinees. The examinees may also have remembered some of the items and thought about them between the two administrations of the test. The examinees may have had different levels of motivation to achieve, or the timing of the test administration with respect to the school year (the beginning versus the end) may have affected their motivation to achieve or their attention span or willingness to exert effort. Some type of differential item functioning or bias in favor of (or against) one or more of the four intervention conditions in the larger study may exist. Only more rigorous analysis can determine whether that difference can be attributed to the intervention.

7.2.3. Test Information

Because the final scales of the two tests shared 16 items out of 17, the shape of the test information curve of the postintervention test was very similar to that of the preintervention test. In the preintervention test and the equated scale, person-ability estimate was related to lower CSEM (i.e., higher test information) when the former ranged between –1.00 and 1.00. In the unequated postintervention test, the person-ability estimate was associated with higher test information and lower CSEM for the person ability estimates between –1.20 and 0.00 on the θ scale.

7.3. External Validity

The K-TEF test will be used as an outcome variable in statistical models designed to estimate the effect of the intervention on educators' MKT. A previous version of the K-TEF test was used in a previous randomized trial (Lewis & Perry, 2017). According to CTT-based scoring methods, the previous version of the test detected a significant difference in 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 it. Likewise, whether the scores on the K-TEF test will significantly predict student learning or moderate the effect of the intervention on student learning is not yet known. These will be the next steps in investigation of the validity argument for the K-TEF.

Because the spring 2017 K-TEF, used as a postintervention measure of educator abilities in a randomized trial of an educational intervention, was identical to the preintervention test with the exception of one item, examinees' preintervention scores should be a reasonably strong predictor of their postintervention scores. The Pearson correlation coefficient for the 228 examinees who participated in both the tests for the θ estimates from the EAP estimator was .67. The two tests were administered at least six months apart, and an intervention took place between them for approximately three-quarters

of the examinees in the sample. We interpret the correlation to indicate the K-TEF to be providing a reasonably stable estimate of educators' abilities.

7.4. Conclusions

On the basis of the sample of 241 educators' responses to items from the spring 2017 field test of the K-TEF test, the 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 how well the test is well-suited for its intended use. Evaluation of the structural validity of the resulting 17-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, but no examinees received a zero score. 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 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.

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

Kang, T., & Petersen, N. S. (2009). *Linking item parameters to a base scale* (ACT Research Report Series No. 2009-2). Iowa City, IA, ACT, Inc.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer Verlag.

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.

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.

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. 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.

Schoen, R. C., Bray, W., Wolfe, C., Nielsen, L., & Tazaz, A. M. (2017). Developing an assessment instrument to measure early elementary teachers' mathematical knowledge for teaching. *The Elementary School Journal, 118*(1), 55–81. https://doi.org/10.1086/692912

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

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

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

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.

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.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17*(2), 253–269.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432–442.

Appendix A. Sources of Assessment Items

Item number	Correct response	Item description	Item original source	Coded qualitatively
Q1A	Yes (1)		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
Q3		Maria needs to swim 6 miles in eight days.	Schoen, 2018	N
Q4	A (1)	Point closest to	Learning Mathematics for Teaching (LMT) [1]	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]		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]		Highway under construction	Zhou, Peverly, & Xin, 2006	N
Q10A[a]			Newton, 2008	N
Q10B[a]			Newton, 2008	N
Q10C[a]			Newton, 2008	N
Q11[a]		Given , how many ropes?	Schifter, 1998	Y
Q12	E (5)	Student representations of	LMT [2]	N
Q13	C (3)	Jim's proportion of program sessions taught	LMT [3]	N
Q14A	2 (No)	Word problem for	LMT [4]	N
Q14B	2 (No)	Word problem for		N
Q14C	1 (Yes)	Word problem for		N
Q14D	1 (Yes)	Word problem for		N
Q15	B (2)	Divide students	LMT [5]	N
Q16	E (5)		LMT [6]	N
Q17	–	Line segment of	Beckmann, 2005	Y
Q18	C (3)	Models to	LMT [7]	N
Q19	--	Connections - measurement and fractions	MCLSG	Y
Q20	C (3)	Fractional part of square is triangle A	LMT [8]	N
Q21	C (3)	Paper frog moving along a line	LMT [9]	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 students to answer "how many	MCLSG	Y
Q24	–	Similarities & differences between fractions & whole numbers	MCLSG	Y
Q25A	2 (No)	Word problem	LMT [10]	N
Q25B	1 (Yes)	Word problem		N
Q25C	2 (No)	Word problem		N
Q25D	1 (Yes)	Word problem		N
Q26	B (2)	Comparing	LMT [11]	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, 2001A-16
[2] Rational Number, Form B-1
[3] Elementary Number Concepts & Operations, Content Knowledge, 2001B-3
[4] Rational Number, Form B-9
[5] Elementary Number Concepts & Operations, Knowledge of Content and Students, 2001A-13
[6] Rational Number, Form A-6
[7] Elementary Number Concepts & Operations, Content Knowledge, 2001B-17
[8] Elementary Number Concepts & Operations, Content Knowledge, 2001B-5
[9] Rational Number, Form A-4
[10] Rational Number, Form A-10
[11] Rational Number, Form B-6

Appendix B. Knowledge for Teaching Elementary Fractions Test

****Appendix B has been redacted for test security. For information about specific items or to request the test for use in your own work, contact the lead author:**
Robert Schoen, rschoen@lsi.fsu.edu.