Development and Initial Field Test of the 2016 K-TEEM (Knowledge for Teaching Early Elementary Mathematics) Test

Robert C. Schoen
Xiaotong Yang
Amanda M. Tazaz
Wendy S. Bray
Kristy Farina

OCTOBER 2019

Research Report No. 2019-01

SECURE VERSION

The research and development reported here were supported by grants to Florida State University from the Institute of Education Sciences, U.S. Department of Education (grant award number R305A120781) and from the Florida Department of Education (grant award number 371-2355B-5C001). The opinions expressed are those of the authors and do not represent views of the institute, the U.S. Department of Education, or the Florida Department of Education.

This work was reviewed and overseen by the Florida State University Institutional Review Board (FWA No. IRB00000446) as HSC number 2015.14936.

Suggested citation: Schoen, R. C., Yang, X., Tazaz, A. M., Bray, W., & Farina, K. (2019). *Development and Initial Field Test of the 2016 K-TEEM (Knowledge for Teaching Early Elementary Mathematics) Test* (Research Report No. 2019-01). Tallahassee, FL: Florida State University. DOI: 10.33009/fsu.1581610055.

Development and Initial Field Test of the 2016 K-TEEM (Knowledge for Teaching Early Elementary Mathematics) Test

Research Report No. 2019-01

Robert C. Schoen
Xiaotong Yang
Amanda M. Tazaz
Wendy S. Bray
Kristy Farina

October 2019

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

Acknowledgements

The successful development and implementation of this assessment involved many experts in mathematics education and many more practicing educators. Some of the key people involved with the development of the test are listed here along with their roles in the endeavor.

Robert Schoen designed the content and format of the test, directed the study, created the scoring criteria, interpreted the results, and coordinated the writing of this report. Xiaotong Yang performed the data analysis for the missing-data analysis, dimensionality analysis, item calibration, and item-response theory-based models. Amanda Tazaz created the test form in the Qualtrics platform, provided technical support for participants, reviewed items, and assisted with recruitment of and communication with examinees. Wendy Bray assisted with the development of the test items and the scoring criteria. Kristy Farina provided support for data management, coordinated assignments for scoring and adjudication of constructed-response items, verified the accuracy of the data, and assisted with description of the sample and scoring criteria. Each of these coauthors also contributed to the writing and editing of this report.

Many additional individuals contributed to the success of the development and initial field test of the 2016 K-TEEM. Some of them are named here. Naomi Iuhasz-Velez and Mark McClure reviewed items and provided feedback on them in the early stages of development. Robert Schoen, Amanda Tazaz, Wendy Bray, Charity Buntin, Kristy Farina, Zachary Champagne, and Claire Riddell participated in the item review and the adjudication meetings. Anne Thistle provided valuable assistance with copy editing. Charity Buntin provided assistance with editing the style and format of the report.

We are especially grateful for the support from the Florida Department of Education, the Institute of Education Sciences, and for the educators who agreed to participate in the study.

Table of Contents

Acknowledgements iv

Executive Summary x

Background x

Content and Structure x

Description of the Sample x

Data Analysis and Scoring xi

Conclusions and Future Directions xi

1. Introduction 1

1.1. Item Development 2

1.2. Features of Testing Platform 3

2. Initial Item Review 4

3. Data and Scoring 6

3.1 Description of the Sample 6

3.2 Missing Response Data 8

3.3. Data Entry and Verification Procedures 8

3.4. Item Scoring 8

3.5. Item Removal 9

4. Dimensionality Analysis 11

5. Classical Test Theory (CTT) Analyses 12

5.1. Distribution of the Observed Test Score 12

5.2. Item Difficulty and Discrimination 12

5.3. Coefficient α and Standard Error of Measurement 14

6. Item Response Theory (IRT) Analyses 15

6.1. Model Description 15

6.2. Item Difficulty and Discrimination 15

6.3. Test Information and Estimated Person Ability 18

7. Discussion and Conclusions 21

References 22

List of Appendices

Appendix A. 2016 K-TEEM Items in Test-Form Order with Scoring Key 23

Appendix B. Administration Instructions at the Start of the Online Assessment 58

Appendix C. Further Specification and Exemplars of Scoring Criteria for Constructed Response Items ... 59

Appendix D. Proportion of Teacher Responses by Item 64

List of Tables

Table 1.1. Test Blueprint for the Original 2016 K-TEEM Test Form and the Final Scale 2
Table 2.1. Test Blueprint for the 2016 K-TEEM Test, Split by Phase in Data Analysis 5
Table 3.1. Self-Reported Characteristics of Teachers Participating in the 2016 K-TEEM Field Test 7
Table 3.2. Missing Response Frequency in the Sample 8
Table 3.3. Item Indexing and Scoring for both Test-Form and Final-Scale Formats 10
Table 5.1. Item Difficulty and Discrimination from CTT Analyses 13
Table 5.2. Distribution of Item Difficulty and Discrimination Estimates for the Items in the Final Scale 14
Table 6.1. Descriptive Statistics of Discrimination Estimates and Difficulty Estimates of Each Item 15
Table 6.2. Parameter Estimates and Standard Errors for Final-Scale Items Modeled Using 2PL 16

List of Figures

Figure 4.1. Parallel analysis scree plot. 11
Figure 5.1. Bar graph depicting the distribution of the observed test score in the final-scale format. 12
Figure 6.1. Item discrimination estimate (a) of each final-scale item. 17
Figure 6.2. Item difficulty estimate (b) of each final-scale item. 17
Figure 6.3. Test information curve and CSEM for the final scale format. 18
Figure 6.4. Person abilities (i.e., θ) estimated by maximum likelihood estimation (MLE). 20
Figure 6.5. Person abilities (i.e., θ) estimated by expected a posteriori (EAP). 20

List of Equations

Equation 1. Standard Error of Measurement (SEM) from CTT Analyses 14

Equation 2. Two-Parameter (2PL) Model 15

Equation 3. Conditional Standard Error of Measurement (CSEM) Given Person Ability 18

Executive Summary

The *2016 Knowledge for Teaching Early Elementary Mathematics* (2016 K-TEEM) test measures teachers' mathematical knowledge for teaching early elementary mathematics. The intended use of the test is twofold. Its first purpose is to estimate the effect of professional-development programs designed for educators who are responsible for teaching mathematics to students in the early elementary grade levels on those educators' mathematical knowledge for teaching mathematics at those grade levels. The second is to support empirical research into the association between teacher mathematical knowledge for teaching and other facets of the teaching and learning process, including teacher beliefs, instructional practice, and student learning.

The purpose of this report is to present results of the first large-scale field test of the 2016 K-TEEM test with 383 practicing educators. It contains information about the development process used for the test, a description of the sample, descriptions of the procedures used for data entry, scoring of responses, and analysis of data, recommended scoring procedures, and findings regarding the distribution of test scores, standard error of measurement, and reliability estimates. This report speaks to the substantive and structural elements of validity (Flake, Pek, & Hehman, 2017). Future work will examine the external validity of the test scores.

Background

The *2016 K-TEEM* is the third version of the K-TEEM (Schoen, Bray, Wolfe, Tazaz, & Nielsen, 2017). At the time of this writing, four versions exist. Each version is named for the first year in which it was subjected to a large-scale field test. More than half of the items on the 2016 K-TEEM were also on the 2014 and 2015 K-TEEM forms. The process of generating and refining new items for the 2016 K-TEEM was the same as that used for the initial development of the 2014 K-TEEM test form. (See Schoen et al., 2017, for further explication of the development process.)

Content and Structure

Within the mathematical knowledge for teaching (MKT) theoretical framework (Ball, Thames, & Phelps, 2008), the 2016 K-TEEM contains items designed to measure teachers' common content knowledge, specialized content knowledge, knowledge of content and students, knowledge of content and teaching, and knowledge of content and curriculum. The 2016 K-TEEM test assesses teachers' MKT in the domains of number, operations, and algebraic thinking. It is not designed to measure teachers' knowledge or abilities in other aspects of mathematics, such as geometry, measurement, or data analysis.

Approximately 22% of the items contributing to the final score are designed to assess knowledge attributed to the domains of *Knowledge of Content and Students* and *Knowledge of Content and Curriculum*, but the emphasis in the 2016 K-TEEM is on content knowledge. Three items use a constructed-response format; the remaining 29 items that contribute to the final scale use a selected-response format.

Description of the Sample

The 2016 K-TEEM was field tested with 383 elementary educators in Florida during spring/summer 2016. The majority of the examinees (n = 311) identified themselves as kindergarten-, first-, or second-grade teachers. Some of the examinees (n = 28) identified themselves as intermediate-grades teachers. The remaining examinees identified themselves as instructional support personnel (e.g., mathematics coaches, intervention specialists).

Approximately two-thirds (65%) of the teachers in the sample had attended at least one year of a professional-development program based on Cognitively Guided Instruction (CGI) before the administration of the 2016 K-TEEM, leaving 137 examinees (35%) who had not yet participated in any part of the CGI program.

Data Analysis and Scoring

The 2016 K-TEEM is composed of 32 items contributing to the final-scale score. Parallel analysis supported the assumption of essential unidimensionality.

We analyzed the data using both classical-test-theory and item-response-theory approaches. According to the first, item-difficulty estimates ranged from .21 to .83, and the item discrimination estimates ranged from .20 to .44. The minimum total raw score (out of 32) was four, and 0.5% of examinees in this sample received a perfect score of 32. Coefficient α and conditional standard error of measurement were found to be .83 and 2.46, respectively, with the field-test sample.

For the second, analyses, we used a two-parameter model for calibration. The distribution of θ scores according to the *expected a* posteriori method with the field-test sample ranged from –2.06 to 2.57. The mean was 0.00 with a standard deviation of 0.91. Because of the perfect scores, the *expected a posteriori* method of estimating the person ability of each examinee is recommended.

Conclusions and Future Directions

The overall difficulty of the 2016 K-TEEM test and the abilities of the educators in the 2016 field-test sample aligned reasonably well, and the reliability estimates appear to be sufficiently high for the intended use of the test. The results described in the present report provide an argument in favor of the substantive and structural aspects of validity (Flake, Pek, & Hehman, 2017). Future validation efforts will determine whether the effects of teacher professional-development interventions can be detected by the 2016 K-TEEM scores and whether the 2016 K-TEEM scores are associated with other factors, such as student learning in mathematics.

1. Introduction

The *2016 Knowledge for Teaching Early Elementary Mathematics* (K-TEEM) test measures teachers' mathematical knowledge for teaching early elementary mathematics (Ball, Thames, & Phelps, 2008; Schoen, Bray, Wolfe, Tazaz, & Nielsen, 2017). Within the mathematical knowledge for teaching (MKT) theoretical framework, the 2016 K-TEEM contains items designed to measure teachers' common content knowledge, specialized content knowledge, knowledge of content and students, knowledge of content and teaching, and knowledge of content and curriculum. The 2016 K-TEEM focuses on the domains of number, operations, and algebraic thinking. It is not designed to measuring teachers' knowledge or abilities in other aspects of mathematics, such as geometry, measurement, or data analysis.

As of this writing, four versions of the K-TEEM exist, each named for the year during which it was first field tested with teachers. The first version was the 2014 K-TEEM (Schoen, Bray, Wolfe, Tazaz, & Nielsen, 2017). The second, the 2015 K-TEEM, was almost identical to the first. The 2016 K-TEEM was the third, and the 2016 K-TEEM the fourth. Although it does include items designed to measure teachers' knowledge of content knowledge, pedagogical content knowledge, and curricular knowledge (Shulman, 1986), the 2016 K-TEEM focuses more on content knowledge than the previous two tests, which included more items designed to measure teachers' pedagogical content knowledge and curricular knowledge.

The 2016 K-TEEM includes 15 items that were also on the 2015 K-TEEM. Another 6 items have been slightly modified (e.g., minor revisions to wording, response order, or number of response options) from the 2015 version. Of the 14 new items, 10 focus on content knowledge (including both common content knowledge and specialized content knowledge).

Table 1.1 shows the test blueprint for the 2016 K-TEEM original test form and the final scale after data recoding. The blueprint shows the categories measured by the test as well as the number of items corresponding to each category. The 2016 K-TEEM was used to measure teachers' MKT at the end of a year-long randomized-controlled trial of a teacher professional-development program called Cognitively Guided Instruction (CGI).

Table 1.1. Test Blueprint for the Original 2016 K-TEEM Test Form and the Final Scale

Category and subcategory	Abbreviation	Number of items: Test form	Number of items: Final scale
Common Content Knowledge	CCK		
Meaning of the Equal Sign and Related Notation	ES	4	4
Meaning of Terms Expression/Equation	EE	2	2
Properties of Operations	PO	4	3
Solve Problems in Many Ways	SMW	3	1
Specialized Content Knowledge	SCK		
Interpreting Student Strategies	ISS	6	6
Connecting Models of Mathematical Ideas	CMMI	3	3
Modeling the Structure of a Word Problem	MSWP	3	3
Knowledge of Content and Students	KCS		
Relative Problem Difficulty	RPD	3	3
Knowledge of Content and Teaching	KCT		
Selecting Word Problems in Service of Specific Instructional Goals	LG	3	3
Knowledge of Content and Curriculum	KCC		
Naming Word Problem Types	NPT	4	4
Total		*35*	*32*

1.1. Item Development

The process of generating and refining new items for the 2016 K-TEEM was the same as that used for the initial development of the 2014 K-TEEM test form (see Schoen et al., 2017). New items were drafted in relation to a target blueprint specifying subcategories of items within MKT subdomains. Draft items were reviewed by experts in mathematics and mathematics education. These experts were asked to provide feedback on what each item was measuring, language clarity, anticipated responses and possible correct responses, and expected level of item difficulty.

After items were revised or eliminated on the basis of this initial round of expert feedback, the remaining items were used in cognitive interviews with six practicing teachers. In the cognitive interviews, teachers were asked to answer each draft item (as if they were taking the test) and to verbalize their thinking processes while and after working on each task. Interviewers also asked the teachers probing questions to gauge further their thinking with respect to the aspects of mathematics and mathematics teaching and learning relevant to each item. The cognitive interviews provided insight regarding how teachers interpreted tasks and response options and what aspects of the items they found confusing. The data collected through cognitive interviews was reviewed by the development team, and items were further revised or eliminated. Items that remained at the end of this process were put into final form and inserted into the Qualtrics (2005–2014) online survey platform.

Responses for items that use a constructed-response format were reviewed by an adjudication committee and/or scored according to rubrics as described below.

1.2. Features of Testing Platform

Like the previous K-TEEM tests, the 2016 K-TEEM test was administered in a Web-based format by means of the Qualtrics (2005–2014) platform. This platform affords a multimedia approach, thereby supporting the use of images and videos in the items. Examinees accessed the test form through a personalized link that was sent by e-mail directly to each individual person. Examinees were validated against a testing database before gaining access to the platform. Information about how to seek technical support was displayed at the bottom of every page and was available throughout the testing window. All questions were formatted on the platform to allow the test to be completed on a computer or a mobile device. All items in the test were displayed one item per page, and a progress bar appeared at the bottom center of the screen.

All questions used forced response; an answer had to be recorded before the examinee was allowed to advance to the next question. After each response was submitted—an active and deliberate action taken by the examinee—the software did not allow the examinee to return to view any previous questions or change any response. If an examinee was unable to complete the assessment in one sitting, the entry validation allowed for ending that session and returning at a later time to continue the test, starting with the first item in the sequence that had not yet been submitted.

2. Initial Item Review

The 2016 K-TEEM test contained 35 numbered items, each assigned an item code reflecting the associated subdomain of MKT. (See Table 2.1.) These items prompted up to 50 responses from the examinee, because three items required multiple responses. Item CCK.SMW.6 provided 6 fields in which examinees could describe up to 6 ways of solving the problem. Item CCK.ES.3 provided 4 blanks for examinees to complete. Appendix A contains the scoring criteria for these two items. In addition, item CCK.EE.1 was a testlet that included 8 subitems. Each subitem was a question that prompted a dichotomous (i.e., true, false) response. Section 3.4 provides more details about the data analysis and scoring decisions related to this item.

The 35 numbered items were coded into 35 dichotomous (i.e., correct, incorrect) variables. The initial test therefore consisted of 32 selected-response items and 3 constructed-response items. Items 4, 20 and 35 were not included in the final scale, for several statistical reasons, which are explained in section 3.4 below.

To clarify the item recoding process here, we labeled the sets of 50, 35, and 32 items the data-entry, test-form, and final-scale formats of the test, respectively. To differentiate test-form and final-scale items, we placed an asterisk after each final-scale item number (See Table 2.1). After item 3 was recoded and items 4, 20, and 35 excluded, 32 items remained to contribute to the final-scale score, including 29 selected-response items and 3 constructed-response items. Table 2.1 provides a blueprint for the test and includes a map of the correspondence among the data-entry, test-form, and final-scale formats.

Table 2.1. Test Blueprint for the 2016 K-TEEM Test, Split by Phase in Data Analysis

Item	Test-form item #	Final-scale item #
Common Content Knowledge		
CCK.ES.3	10	9*
CCK.ES.5	30	28*
CCK.ES.7	22	20*
CCK.ES.2	25	23*
CCK.EE.1	3	3*
CCK.EE.2	7	6*
CCK.PO.2	29	27*
CCK.PO.7	14	13*
CCK.PO.8	20	
CCK.PO.9	33	31*
CCK.SMW.6	5	4*
CCK.SMW.7	35	
CCK.SMW.8	4	
Specialized Content Knowledge		
SCK.CCMI.3	24	22*
SCK.CMMI.2	17	16*
SCK.CMMI.4	19	18*
SCK.ISS.1	11	10*
SCK.ISS.2	9	8*
SCK.ISS.3	27	25*
SCK.ISS.4	15	14*
SCK.ISS.5	32	30*
SCK.ISS.6	31	29*
SCK.MSWP.1	18	17*
SCK.MSWP.2	26	24*
SCK.MSWP.3	23	21*
Knowledge of Content and Students		
KCS.RPD.4	28	26*
KCS.RPD.5	34	32*
KCS.RPD.6	1	1*
Knowledge of Content and Teaching		
KCT.LG.1	2	2*
KCT.LG.2	13	12*
KCT.LG.5	21	19*
Knowledge of Content and Curriculum		
KCC.NPT.1	16	15*
KCC.NPT.12	8	7*
KCC.NPT.14	6	5*
KCC.NPT.15	12	11*

3. Data and Scoring

3.1 Description of the Sample

The web-based 2016 K-TEEM test was completed by 387 practicing teachers in spring 2016. The items on the 2016 K-TEEM test and scoring key are provided in Appendix A. Administration instructions accompanying the test are provided in Appendix B. Additional specification of the scoring criteria for two constructed-response items are provided in Appendix C. Administration of the tests occurred during a period spanning April 26–August 6, 2016. Three hundred forty-eight (90%) of the teachers in this sample completed the test between April 26 and May 16, 2016, and 39 completed it between May 16 and August 6, 2016.

Approximately two-thirds (65%) of the teachers in the sample had attended at least one year of a professional-development program based on Cognitively Guided Instruction (CGI) before completing the 2016 K-TEEM. The CGI program offers up to three years of training. In the 2016 K-TEEM field-test sample, 205 of the participating teachers (53%) had completed exactly year one of the program, 26 (7%) had completed two years, and 24 (6%) had completed all three years, leaving 137 examinees (35%) who had not participated in any part of the CGI program at the time that they completed the test.

Table 3.1 shows self-reported characteristics of teachers participating in the 2016 K-TEEM field test. The average number of years of teaching experience among the teachers in the sample was 12.10 (SD = 8.67). The minimum number of years of teaching experience reported was 0, and the maximum was 40. Almost all (95%) of the participants in the sample identified themselves as female. The sample consisted mostly of classroom teachers (88%), and the remaining participants identified themselves as filling instructional support roles such as math coach, interventionist, or resource staff. The sample represents 12 school districts, spanning the full geographic range of the state and including urban, suburban, and rural areas.

Table 3.1. Self-Reported Characteristics of Teachers (n = 387) Participating in the 2016 K-TEEM Field Test

	n	Proportion
Gender		
Male	13	.03
Female	369	.95
Decline to answer	3	.01
Unknown	2	.01
Race		
American Indian	4	.01
Asian	7	.02
Black	49	.13
Multiracial	6	.02
White	303	.78
Unknown	3	.01
Decline to answer	15	.04
Hispanic		
Hispanic	66	.17
Not Hispanic	309	.80
Decline to answer	10	.02
Unknown	2	.01
Grade role		
K	60	.16
1	130	.34
2	121	.31
3	17	.04
4	7	.02
5	4	.01
Other instructional support[a]	46	.12
Unknown	2	.01
Years of teaching experience		
Three or fewer	46	.12
Four or more	339	.88
Unknown	2	.01
Highest degree earned		
Associate's degree	2	.05
Bachelor's degree	244	.63
Master's degree	128	.33
Professional diploma	8	.02
Professional degree	3	.01
Unknown	2	.01

Note. Proportions may not sum to 1 because of rounding.

[a] The Other Instructional Support category includes 12 teachers who were coded as "other," 30 teachers who were specifically coded as "Support," and 4 teachers who were coded as "Multiple Grade Levels."

3.2 Missing Response Data

Four examinees did not provide complete responses in the test. The forced-response feature prevented examinees from skipping items, but these four examinees ended the test after completing one or more items. We decided to exclude these four teachers from the data analysis, because they had a response rate lower than 50%. Table 3.2 shows the frequency and percentage of missing responses in the sample. After the four incomplete responses were excluded, the analytic sample included 383 educators.

Table 3.2. Missing Response Frequency in the Sample

Number of Missing response(s)	Frequency	%	Cumulative %
0.00	383	98.97	98.97
23.00	2†	.52	99.49
31.00	1†	.26	99.75
41.00	1†	.26	100.00
Total	387	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).

3.3. Data Entry and Verification Procedures

Teachers accessed test items through a personalized link to an online questionnaire hosted within Qualtrics. Teachers entered their own responses through a combination of text-entry boxes and point-and-click, multiple-choice responses. The response data were exported from Qualtrics to a CSV file, which was then transferred to the SPSS platform (IBM Corp., 2017) for scoring. Selected-response items were scored by machine within the SPSS platform.

The responses to the constructed-response items were exported to Microsoft Excel and scored by trained members of the scoring committee using the criteria described in Appendix A and further specified in Appendix C. Raters entered their scores into Excel, and those ratings were merged back into the SPSS file. The result was a file with dichotomous (correct/incorrect) variables. This “raw-score” file was then used for subsequent analysis.

After the four responses to item CCK.ES.3 were reviewed by the adjudication committee, the responses to this item were also scored by machine within SPSS, because the review of responses determined that one, and only one, set of four responses in the empirical data was correct.

3.4. Item Scoring

As explained above, the 50 data-entry variables were recoded into 35 test-form variables, each representing a response that was judged to be either correct or incorrect. The decrease from recoding of the multiple responses to certain items into single responses. Examinees’ responses to the final set of 32 items, including the item-level percentage-correct values, are provided in Appendix D.

Initially, item 3 had 8 subitems, and we decided to exclude three of them (c, e, and h), from item scoring for these reasons: First, subitems e and h are inequalities, so the content review committee suggested

removal in order to focus the types of expressions included in the item on those with equals signs and those without equals signs. Second, subitems c and f are redundant in that they are similar types of equations, they may be relatively trivial for the teaching population, and c is easier than f.

We considered coding the 5 remaining subitems either polytomously (that is, to give the overall item a score of 0 to 5 depending on the number subitems answered correctly) or dichotomously (that is, counting the item correct only if all five subitems were answered correctly and otherwise incorrect). We chose to code item 3 dichotomously, first because coding it polytomously would make it count as a relatively large portion of the total test score and second because of the effect on the items item-rest correlation. Coded dichotomously, its item-rest correlation would be .30, but coded polytomously, its item-rest correlation would be .27. Coding item 3 dichotomously yielded 35 items, each coded dichotomously.

3.5. Item Removal

After removing several of the sub-items from item 3, we decided to exclude items 4, 20,and 35 from the final scale on the basis of the following statistical-analysis results. First, according to the CTT results, these three items had low or negative discrimination estimates: .04, –.02 and –.03 respectively. Second, on the basis of the polychoric correlations, these three items were negatively correlated with a large number of the other items. Table 3.3 shows item indexing and scoring of both test-form and final-scale items.

Table 3.3. Item Indexing and Scoring for both Test-Form and Final-Scale Formats

Item-bank code	Test-form item #	Test-form item score	Final-scale item #	Final-scale item score
KCS.RPD.6	1	0, 1	1*	0, 1
KCT.LG.1	2	0, 1	2*	0, 1
CCK.EE.1	3	0, 1	3*	0, 1
CCK.SMW.8	4	0, 1		
CCK.SMW.6	5	0, 1	4*	0, 1
SCK.NPT.14	6	0, 1	5*	0, 1
CCK.EE.2	7	0, 1	6*	0, 1
SCK.NPT.12	8	0, 1	7*	0, 1
SCK.ISS.2	9	0, 1	8*	0, 1
CCK.ES.3	10	0, 1	9*	0, 1
SCK.ISS.1	11	0, 1	10*	0, 1
SCK.NPT.15	12	0, 1	11*	0, 1
KCT.LG.2	13	0, 1	12*	0, 1
CCK.PO.7	14	0, 1	13*	0, 1
SCK.ISS.4	15	0, 1	14*	0, 1
SCK.NPT.1	16	0, 1	15*	0, 1
SCK.CMMI.2	17	0, 1	16*	0, 1
SCK.MSWP.1	18	0, 1	17*	0, 1
SCK.CMMI.4	19	0, 1	18*	0, 1
CCK.PO.8	20	0, 1		
KCT.LG.5	21	0, 1	19*	0, 1
CCK.ES.7	22	0, 1	20*	0, 1
SCK.MSWP.3	23	0, 1	21*	0, 1
SCK.CCMI.3	24	0, 1	22*	0, 1
CCK.ES.2	25	0, 1	23*	0, 1
SCK.MSWP.2	26	0, 1	24*	0, 1
SCK.ISS.3	27	0, 1	25*	0, 1
KCS.RPD.4	28	0, 1	26*	0, 1
CCK.PO.2	29	0, 1	27*	0, 1
CCK.ES.5	30	0, 1	28*	0, 1
SCK.ISS.6	31	0, 1	29*	0, 1
SCK.ISS.5	32	0, 1	30*	0, 1
CCK.PO.9	33	0, 1	31*	0, 1
KCS.RPD.5	34	0, 1	32*	0, 1
CCK.SMW.7	35	0, 1		

Note. Test-form Item # = the item index from the original test; Final-scale item # = the newly generated item number after excluding items 4, 20 and 35 (we differentiated test-form and final-scale item index by adding * to the final-scale item number).

4. Dimensionality Analysis

Parallel analysis (PA) is a procedure that examines the number of constructs in the data and is considered superior to rule-of-thumb procedures (Wood, Tataryn, & Gorsuch, 1996; Zwick & Velicer, 1982, 1986) such as Kaiser's rule (Kaiser, 1960). After item scoring, we conducted parallel analysis (PA) to examine the dimensionality of the test. The *psych* (Revelle, 2019) program in R 3.6.1 (R Core Team, 2019) was used to perform the analysis.

Figure 4.1 shows the results of the PA. The vertical axis in the figure represents the eigenvalues of principal components, and the horizontal axis represents the number of components. The red dot is for the principal components from the actual data, and the white dot is for those from the resampled data. The number of components from the actual data above the line with white dots indicates the number of dimensions in the data. The confidence intervals for the resampled data were taken into consideration when making the decision. The results suggested that the test was essentially unidimensional.

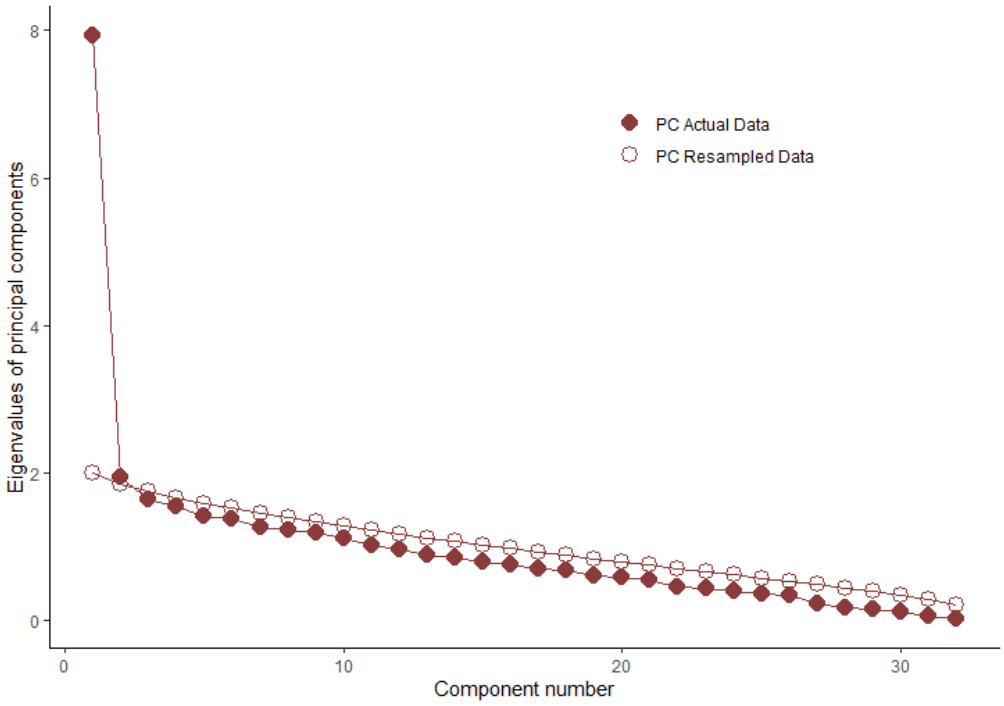


Figure 4.1. Parallel analysis scree plot.

5. Classical Test Theory (CTT) Analyses

After checking the dimensionality of the test, we conducted the classical test theory (CTT) analyses using SPSS 25.0 (IBM corp., 2017).

5.1. Distribution of the Observed Test Score

Figure 5.1 shows the bar graph depicting the distribution of the observed total test score. The total test score for the final scale could have a minimum of 0 and a maximum of 32. The minimum observed score was 4, and the maximum was 32. Two teachers scored 32. The mean of the total test score was 18.03 with a standard deviation of 5.97. The median of the total test score was 18.00. The sample size for these analyses was 383.

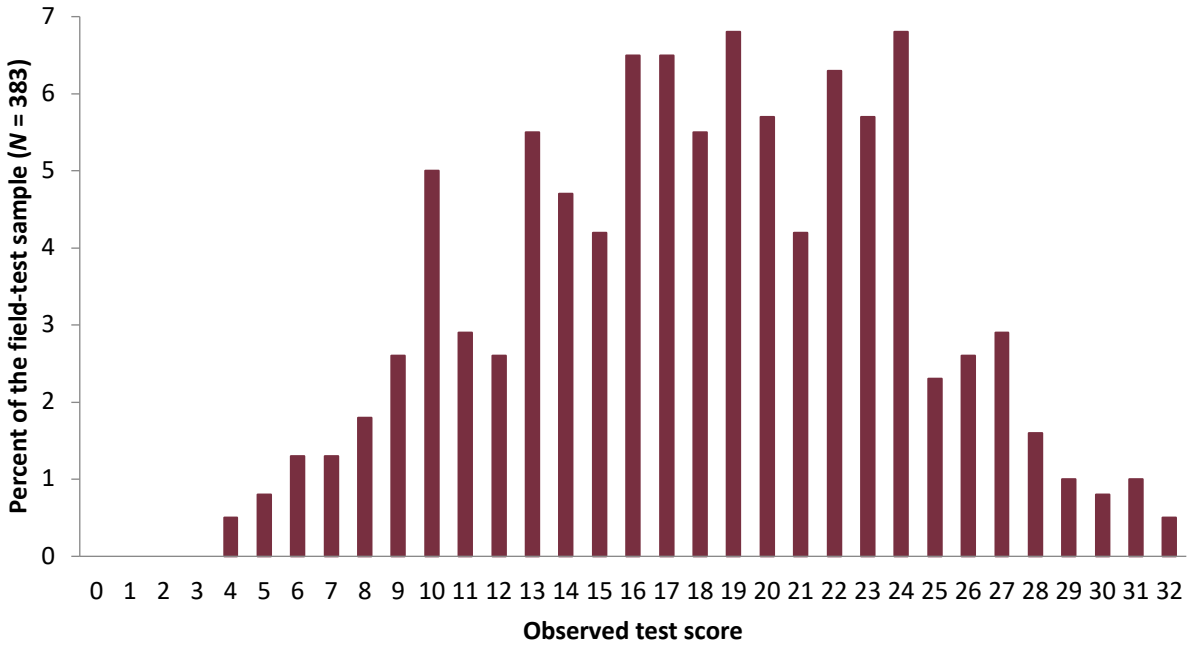


Figure 5.1. Bar graph depicting the distribution of the observed test score in the final-scale format.

5.2. Item Difficulty and Discrimination

We calculated the item difficulty and item discrimination estimates by a CTT approach. Because all the items were dichotomously coded, the item difficulty estimates of each item were calculated as the proportions of correct answers for each item, which were equal to the item means. Table 5.1 shows the descriptive statistics, item difficulty, and item discrimination estimates of each item. Table 5.2 shows the distribution of the CTT-based difficulty estimates and item-rest correlations for the items in the final scale. The item difficulty estimates ranged from .21 (item 4*) to .83 (item 30*). Item discrimination estimates were calculated as the item-rest correlation coefficients (i.e., corrected item-total correlation coefficients) of each item. The item discrimination estimates ranged from .20 (item 22* and item 23*) to .44 (item 8* and item 21*).

Table 5.1. Item Difficulty and Discrimination from CTT Analyses

Item-bank code	Final-scale item #	Scoring	Mean	St. dev.	Item-rest *r*
KCS.RPD.6	1*	0, 1	.80	0.40	.27
KCT.LG.1	2*	0, 1	.74	0.44	.23
CCK.EE.1	3*	0, 1	.61	0.49	.30
CCK.SMW.6	4*	0, 1	.21	0.41	.40
SCK.NPT.14	5*	0, 1	.38	0.49	.26
CCK.EE.2	6*	0, 1	.57	0.50	.38
SCK.NPT.12	7*	0, 1	.80	0.40	.32
SCK.ISS.2	8*	0, 1	.59	0.49	.44
CCK.ES.3	9*	0, 1	.49	0.50	.39
SCK.ISS.1	10*	0, 1	.41	0.49	.23
SCK.NPT.15	11*	0, 1	.52	0.50	.32
KCT.LG.2	12*	0, 1	.61	0.49	.39
CCK.PO.7	13*	0, 1	.56	0.50	.33
SCK.ISS.4	14*	0, 1	.54	0.50	.29
SCK.NPT.1	15*	0, 1	.56	0.50	.35
SCK.CMMI.2	16*	0, 1	.44	0.50	.36
SCK.MSWP.1	17*	0, 1	.64	0.48	.40
SCK.CMMI.4	18*	0, 1	.31	0.46	.30
KCT.LG.5	19*	0, 1	.45	0.50	.23
CCK.ES.7	20*	0, 1	.29	0.45	.36
SCK.MSWP.3	21*	0, 1	.67	0.47	.44
SCK.CCMI.3	22*	0, 1	.56	0.50	.20
CCK.ES.2	23*	0, 1	.43	0.50	.20
SCK.MSWP.2	24*	0, 1	.69	0.46	.40
SCK.ISS.3	25*	0, 1	.42	0.49	.30
KCS.RPD.4	26*	0, 1	.65	0.48	.37
CCK.PO.2	27*	0, 1	.69	0.46	.28
CCK.ES.5	28*	0, 1	.67	0.47	.37
SCK.ISS.6	29*	0, 1	.55	0.50	.33
SCK.ISS.5	30*	0, 1	.83	0.38	.37
CCK.PO.9	31*	0, 1	.74	0.44	.30
KCS.RPD.5	32*	0, 1	.63	0.48	.34

Note. Final-scale item # = the newly generated item number after item recoding (we differentiated test-form and final-scale item index by adding * to the final-scale item number); *M* = 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 Item Difficulty and Discrimination Estimates for the Items in the Final Scale

Value	Number of items
P-value	
>.90	0
.80–.89	3
.70–.79	2
.60–.69	9
.50–.59	8
.40–.49	6
.30–.39	2
.20–.29	2
.10–.19	0
<.09	0
Mean	0.56
Median	0.57
Standard Deviation	0.15
Item-rest correlation	
.80–1.0	0
.60–.79	0
.40–.59	5
.20–.39	27
0.0–.20	0
Mean	0.33
Median	0.33
Standard Deviation	0.07

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

5.3. Coefficient α and Standard Error of Measurement

The coefficient α (Cronbach, 1951) of the test was .83. We subsequently calculated the standard error of measurement (SEM) of the test. The scale variance was 35.64. According to Equation 1, SEM was calculated to be 2.46, where σ^2 is the test variance, and ρ_{XX} is the coefficient α of the test.

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

6. Item Response Theory (IRT) Analyses

6.1. Model Description

We used flexMIRT 3.5 (Cai, 2017) to perform the IRT analyses. The test included 32 items, and all the items were coded dichotomously, as described in above. Although 28 of the items were multiple-choice, we did not use the three-parameter model to adjust for guessing, because the sample size was 383. According to de Ayala (2009), sample sizes exceed 1000 for three-parameter models to be used in IRT calibrations. We therefore used a 2PL model.

Results of flexMIRT indicated that successful convergence was reached in the computation, and the value of –2loglikelihood was 14540.10. The formula of the 2PL model is shown in Equation 2 according to the parameterization of de Ayala (2009).

$$P_j(\Theta) = \frac{\exp\left[a_j(\Theta - b_j)\right]}{1+\exp\left[a_j(\Theta - b_j)\right]}, \quad (2)$$

where a_j is the discrimination index of item *j* (*j = 1, 2,…,J*), b_j is the difficulty index of item *j*, P_j is the probability of correct answer, Θ is the person ability.

6.2. Item Difficulty and Discrimination

Table 6.1 shows the descriptive statistics of the discrimination estimates and the difficulty estimates of each item. The mean of the item discrimination estimates was 0.92 with a standard deviation of 0.26. The mean of the item difficulty estimates was –0.34 with a standard deviation of 0.86. Table 6.2 presents the parameter estimates for each item based on the 2PL model. Figures 6.1 and 6.2 display the item discrimination and item difficulty estimates of each item. The item discrimination estimates ranged from 0.45 (item 23) to 1.39 (item 4). The item difficulty estimates ranged from –1.91 (item 1) to 1.28 (item 4). Ten items had *b* values above 0.00, and 22 items had *b* values below 0.00.

Table 6.1. Descriptive Statistics of Discrimination Estimates and Difficulty Estimates of Each Item

	Mean	St. dev.	Minimum	Maximum	Skewness	Kurtosis
a	0.92	0.26	0.45	1.39	0.17	–0.56
b	–0.34	0.86	–1.91	1.28	0.09	–0.55

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

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

Item-bank code	Final-scale item #	*a*	*s.e.*	*b*	*s.e.*
KCS.RPD.6	1*	0.80	0.19	–1.91	0.42
KCT.LG.1	2*	0.62	0.17	–1.87	0.51
CCK.EE.1	3*	0.71	0.17	–0.71	0.23
CCK.SMW.6	4*	1.39	0.27	1.28	0.21
SCK.NPT.14	5*	0.69	0.16	0.81	0.24
CCK.EE.2	6*	1.02	0.18	–0.35	0.14
SCK.NPT.12	7*	1.07	0.23	–1.57	0.29
SCK.ISS.2	8*	1.31	0.21	–0.38	0.12
CCK.ES.3	9*	1.04	0.19	0.02	0.14
SCK.ISS.1	10*	0.58	0.15	0.67	0.27
SCK.NPT.15	11*	0.83	0.17	–0.13	0.16
KCT.LG.2	12*	1.10	0.18	–0.51	0.14
CCK.PO.7	13*	0.85	0.16	–0.33	0.16
SCK.ISS.4	14*	0.76	0.17	–0.22	0.17
SCK.NPT.1	15*	0.95	0.17	–0.32	0.15
SCK.CMMI.2	16*	0.94	0.18	0.29	0.15
SCK.MSWP.1	17*	1.21	0.21	–0.62	0.14
SCK.CMMI.4	18*	0.83	0.17	1.11	0.26
KCT.LG.5	19*	0.54	0.14	0.42	0.25
CCK.ES.7	20*	1.00	0.18	1.07	0.21
SCK.MSWP.3	21*	1.36	0.23	–0.71	0.14
SCK.CCMI.3	22*	0.46	0.15	–0.56	0.30
CCK.ES.2	23*	0.45	0.14	0.65	0.32
SCK.MSWP.2	24*	1.27	0.22	–0.82	0.16
SCK.ISS.3	25*	0.75	0.15	0.49	0.19
KCS.RPD.4	26*	1.01	0.19	–0.74	0.17
CCK.PO.2	27*	0.74	0.17	–1.24	0.31
CCK.ES.5	28*	1.01	0.18	–0.83	0.18
SCK.ISS.6	29*	0.85	0.17	–0.27	0.16
SCK.ISS.5	30*	1.37	0.26	–1.51	0.24
CCK.PO.9	31*	0.85	0.19	–1.39	0.29
KCS.RPD.5	32*	0.93	0.19	–0.68	0.18

Note. Final-Scale Item # = the newly generated item number after item recoding; *a* = item discrimination index; *b* = item difficulty index; s.e. = standard error.

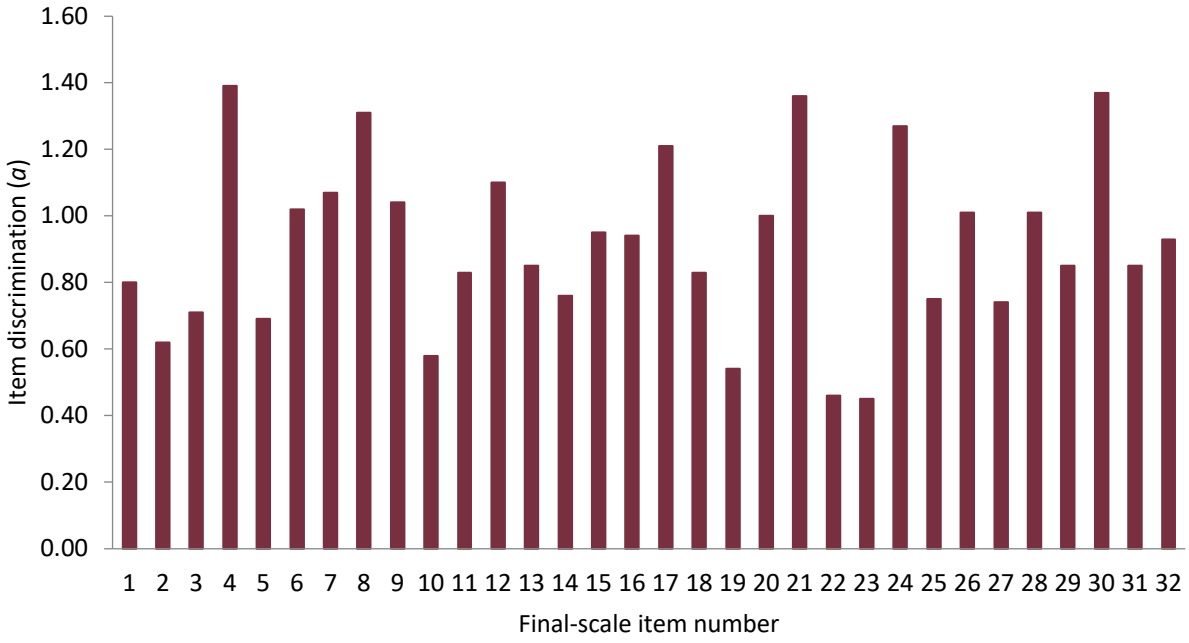


Figure 6.1. Item discrimination estimate (a) of each final-scale item.

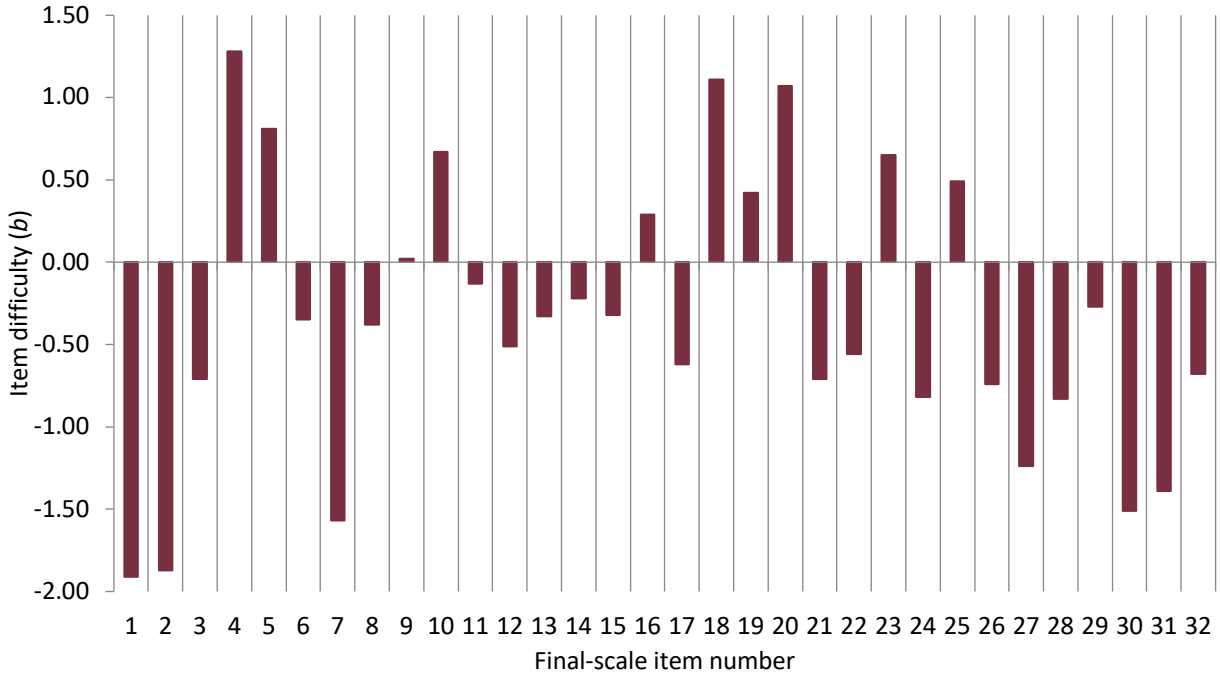


Figure 6.2. Item difficulty estimate (b) of each final-scale item.

6.3. Test Information and Estimated Person Ability

Equation 3 is the formula showing the relationship between the test information and the conditional standard error of measurement (CSEM), where I is the test information function for a given person ability, and θ is the person ability. The formula used to calculate the CSEM was in accordance with the recommendations made by de Ayala (2009).

$$CSEM(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (3)$$

Figure 6.3 shows the relationship between the test information curve and CSEM of the test given person ability estimates. According to Figure 6.3, the person ability (i.e., θ) estimates between –0.80 to –0.40 were associated with the largest test information and the lowest CSEM. Furthermore, the person ability estimates were related to low CSEM (i.e., high accuracy of person ability estimation) when they ranged between –1.60 and 0.80, and they were related to high CSEM (i.e., low accuracy of person ability estimation) when they were smaller than –2.40 or larger than 1.60.

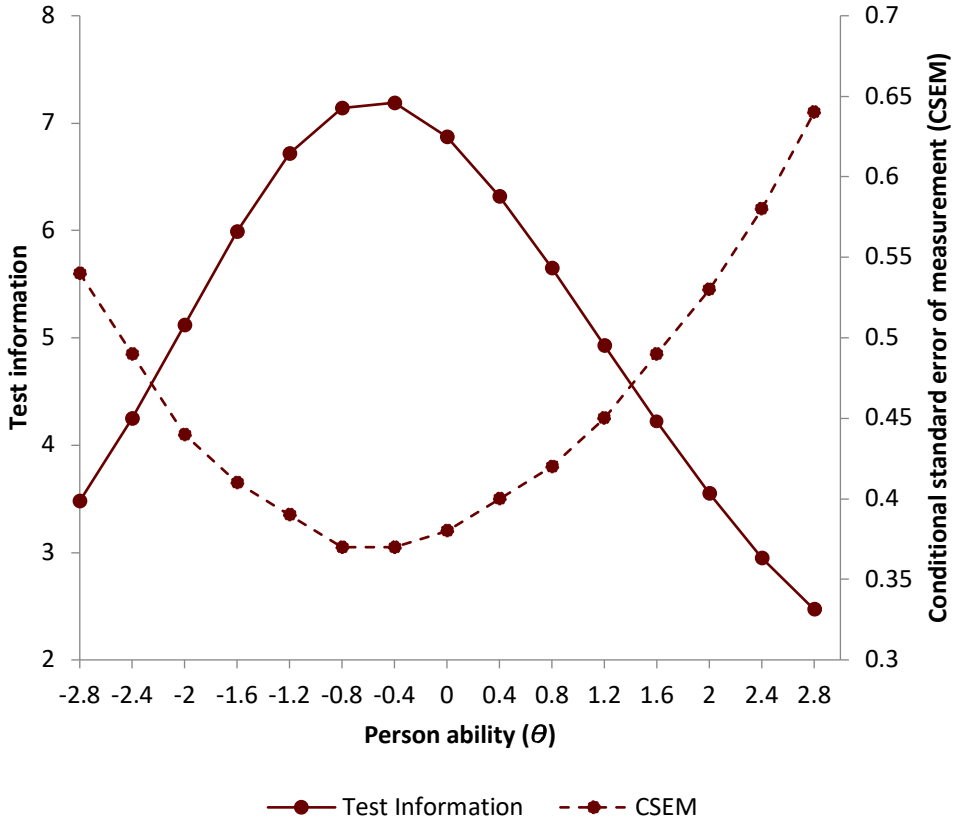


Figure 6.3. Test information curve and CSEM for the final scale format.

We first used maximum likelihood estimation (MLE) to estimate the latent person ability of each individual. Figure 6.4 shows the distribution of person-ability estimation by this method. Note that the spikes at the higher end of the horizontal axis of the distribution curve were a result of the two examinees with perfect scores, whose MLE estimates were not available.

We also used the *expected a posteriori* (EAP) method to estimate the person ability of each individual. Figure 6.5 shows the distribution of person ability estimation by this method. Estimates of θ ranged from –2.06 to 2.57. The mean was 0.00 with a standard deviation of 0.91. The skewness and the kurtosis estimates were 0.17 and –0.15, respectively.

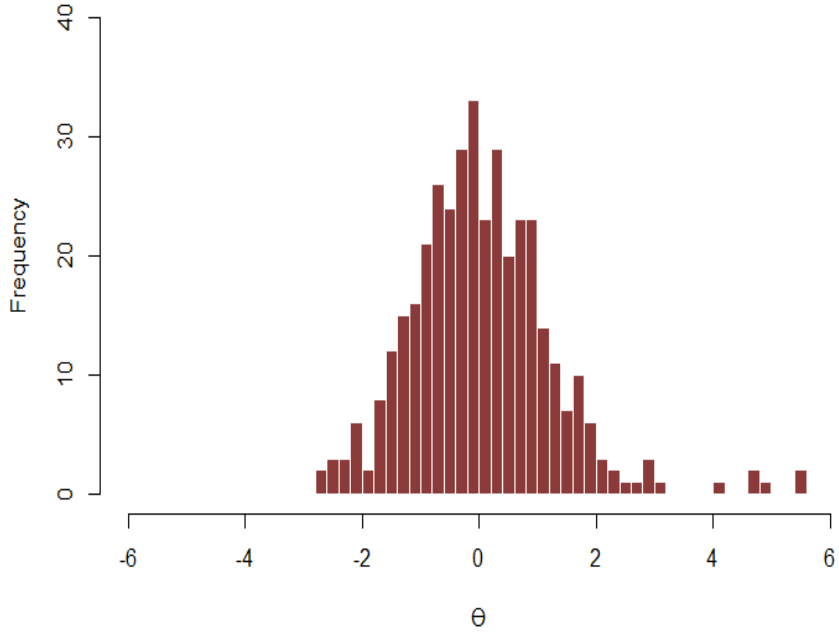


Figure 6.4. Person abilities (i.e., θ) estimated by maximum likelihood estimation (MLE).

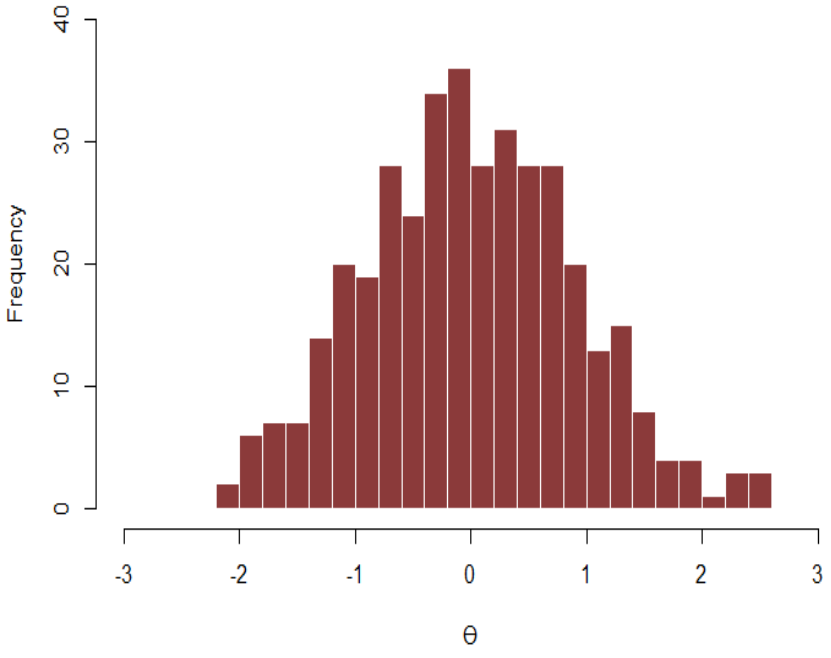


Figure 6.5. Person abilities (i.e., θ) estimated by expected a posteriori (EAP).

7. Discussion and Conclusions

The 2016 K-TEEM test form measures teacher MKT with an emphasis on common content knowledge and specialized content knowledge. Approximately 22% of the items contributing to the final score are designed to assess knowledge attributed to the domains of *knowledge of content and students* and *knowledge of content and curriculum*, but the emphasis in the 2016 K-TEEM is on content knowledge. Three items used a constructed-response format, the remaining 29 items that contribute to the final scale used a selected-response format.

More than half of the items on the 2016 K-TEEM were also on the 2014 and 2015 K-TEEM forms (Schoen, Bray, Wolfe, Tazaz, & Nielsen, 2017). The new items were created through the same development process used for the original items, which included review by content experts and cognitive interviews with practicing, primary-grades teachers.

The sample size with complete data for the 2016 field test was 383. The majority of the examinees (n = 311) identified themselves as kindergarten-, first-, or second-grade teachers. Some of the examinees identified themselves as intermediate-grades teachers (n = 28). The remainder identified themselves as instructional support personnel (e.g., mathematics coaches, intervention specialists).

Results of parallel analysis suggested that the 2016 K-TEEM test was essentially unidimensional. We analyzed the data by both CTT and IRT approaches.

According to the CTT results, the item-difficulty estimates ranged from .21 to .83, and the item discrimination estimates from .20 to .44. Coefficient α was computed to be .83, and the standard error of measurement to be 2.46.

For the IRT analyses, although many test items were multiple-choice, we used a 2PL model for IRT calibration. The resulting item-discrimination estimates ranged from 0.45 to 1.39, and those for item difficulty from –1.90 to 1.28. Ten items had difficulty estimates above 0.00 and 22 below. Person ability (i.e., θ) estimates between –0.80 to –0.40 were associated with the largest test information and the lowest CSEM. Furthermore, the person ability estimates were related to low CSEM (i.e., high accuracy of person ability estimation) when they ranged between –1.60 and 0.80 and to high CSEM (i.e., low accuracy of person ability estimation) when they were smaller than –2.40 or larger than 1.60.

Because the sample included two perfect scores, the EAP method of θ estimation is recommended. The θ estimation by the EAP method ranged from –2.06 to 2.57. The mean was 0.00 with a standard deviation of 0.91. The skewness and the kurtosis estimates were 0.17 and –0.15, respectively.

Future versions of this test should include several additional high-difficulty items in order to discriminate among teachers with the highest ability levels, especially if the test is used with educators who may be likely to have higher-than-average levels of mathematical knowledge for teaching. Nevertheless, the overall difficulty of the 2016 K-TEEM test and the abilities of the educators in the 2016 field-test sample aligned reasonably well, and the reliability estimates appear to be sufficiently high for the intended use of the test.

Future validation efforts will involve analyses to determine whether the effects of teacher professional-development interventions can be detected by the 2016 K-TEEM scores and whether the 2016 K-TEEM scores are associated with student learning in mathematics.

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.

Cai, L. (2017). flexMIRT 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.

IBM Corp. (2017). IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 1–9.

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

Qualtrics. (2005–2014). *Qualtrics software, Version April–May 2014*. Provo, UT: Author.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.9.12. https://CRAN.R-project.org/package=psych.

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

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

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1*(4), 354–365.

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

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

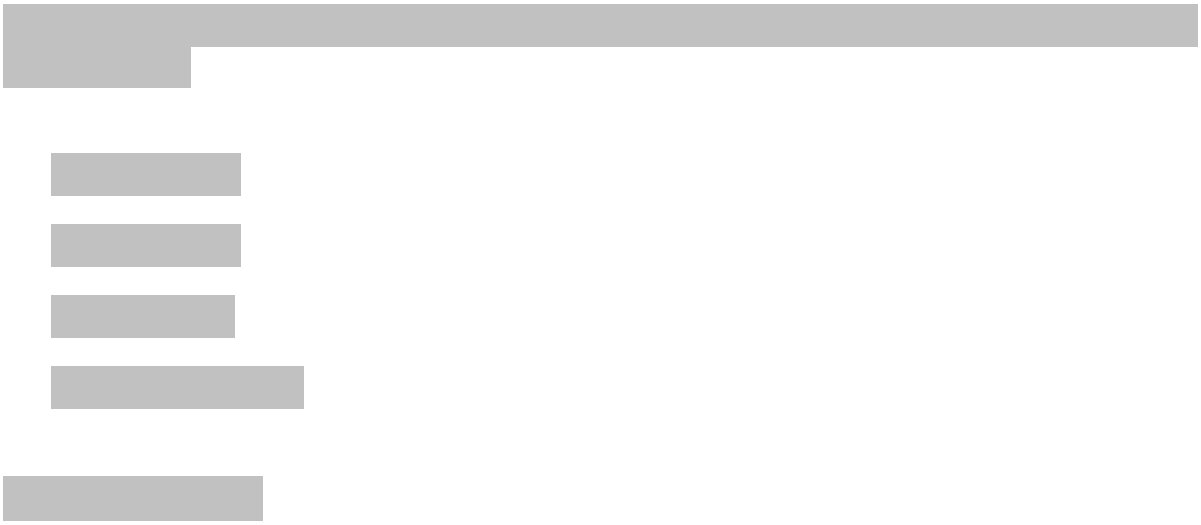
Appendix A. 2016 K-TEEM Items in Test-Form Order with Scoring Key

Test-form item #1
KCS.RPD.6

Test-form item #2
KCT.LG.1

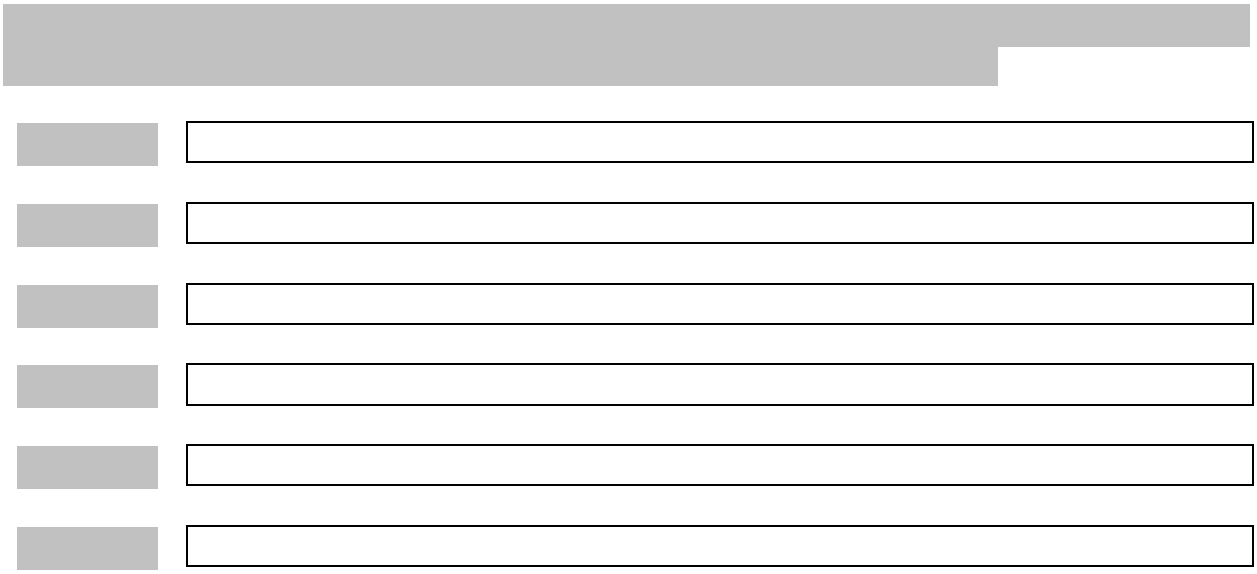
Test-form item #3
CCK.EE.1(a-h)

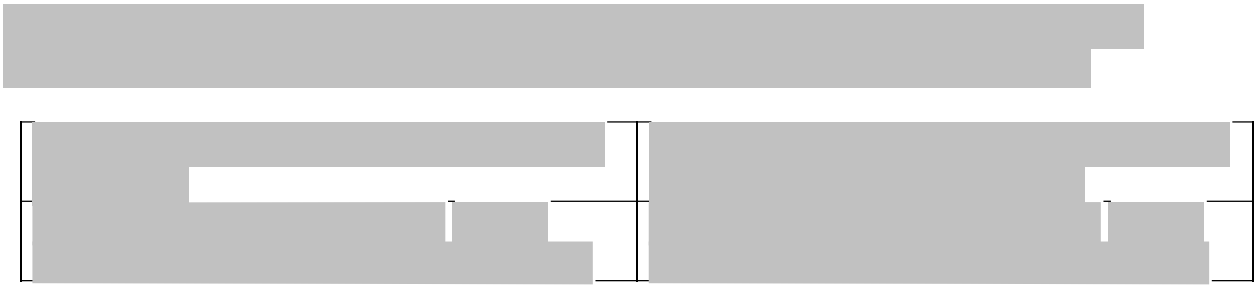
Test-form item #4
CCK.SMW.8



Test-form item #5
CCK.SMW.6

Describe a variety of different strategies that third grade students might use to correctly solve





Test-form item #6
SCK.NPT.14

Test-form item #7
CCK.EE.2

Test-form item #8
SCK.NPT.12

Test-form item #9
SCK.ISS.2

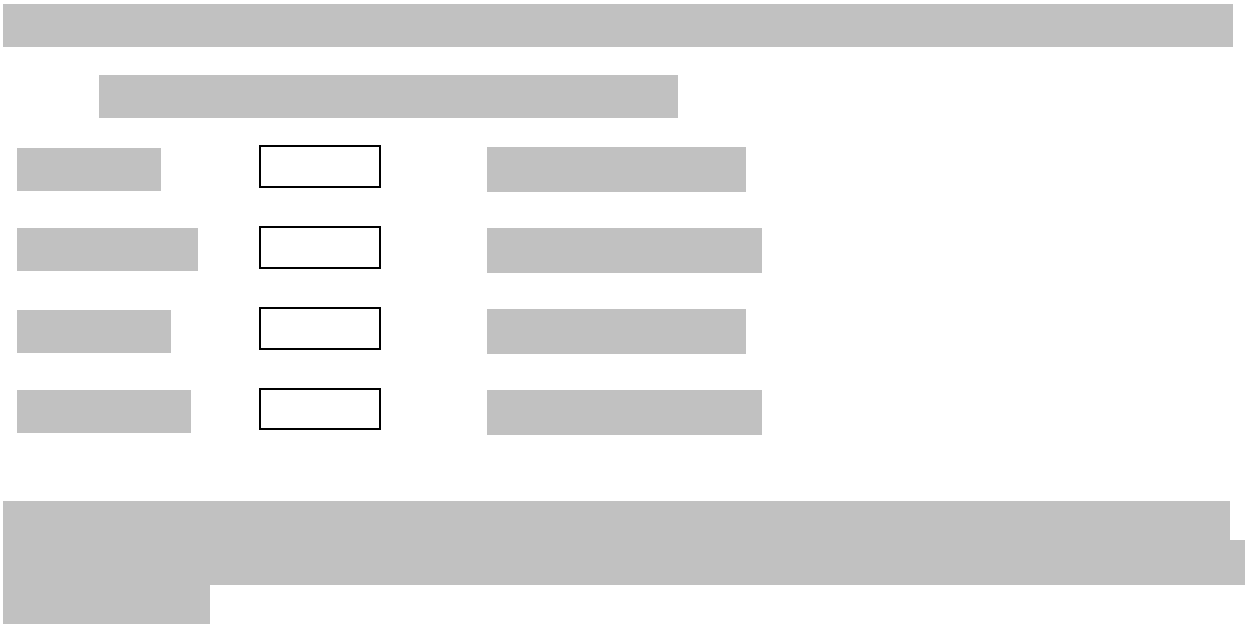
[redacted]

[redacted]

[redacted]

[redacted]	[redacted] .

Test-form item #10
CCK.ES.3

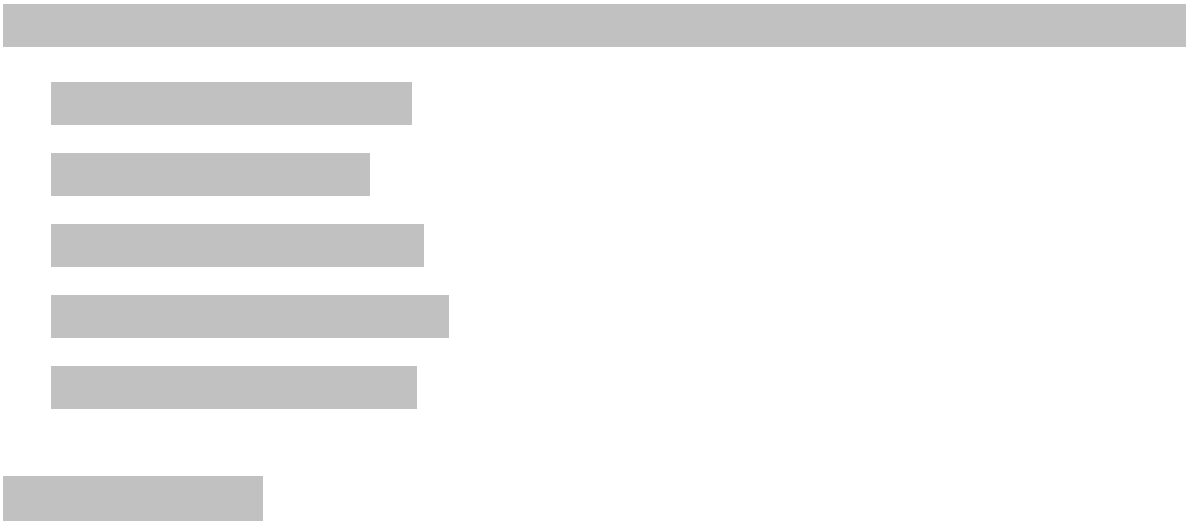


Test-form item #11
SCK.ISS.1

Test-form item #12
SCK.NPT.15

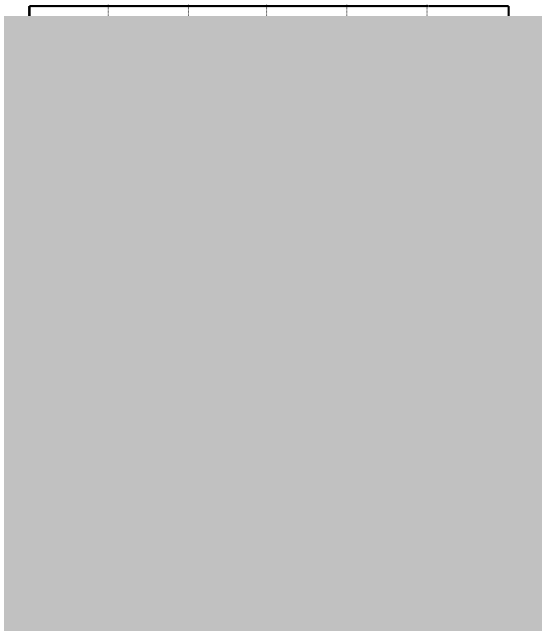
Test-form item #13
KCT.LG.2

Test-form item #14
CCK.PO.7 (Follow-up to KCT.LG.2)



Test-form item #15
SCK.ISS.4

Test-form item #16
SCK.NPT.1



Test-form item #17
SCK.CMMI.2

Test-form item #18
SCK.MSWP.1

Test-form item #19
SCK.CMMI.4

Test-form item #20
CCK.PO.8

Test-form item #21
KCT.LG.5

Test-form item #22
CCK.ES.7

Test-form item #23
SCK.MSWP.3

Test-form item #24
SCK.CCMI.3

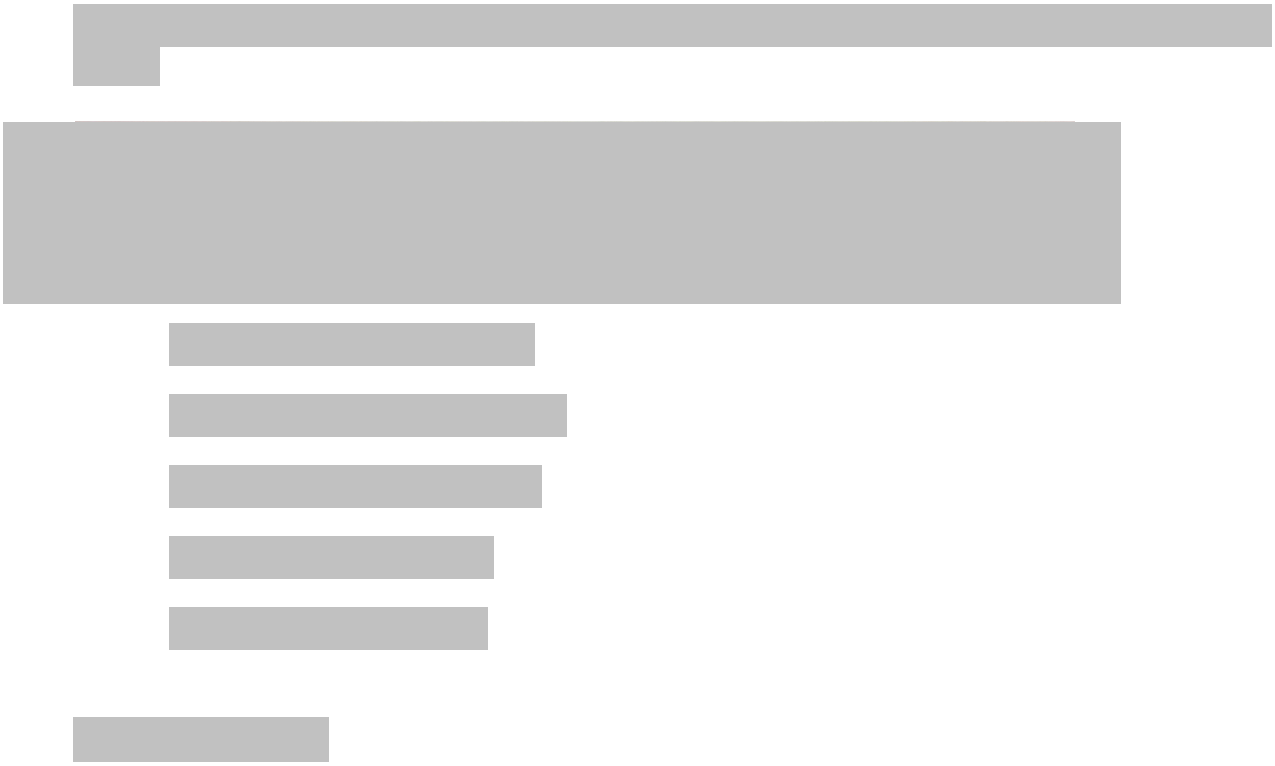
Test-form item #25
CCK.ES.2

Test-form item #26
SCK.MSWP.2

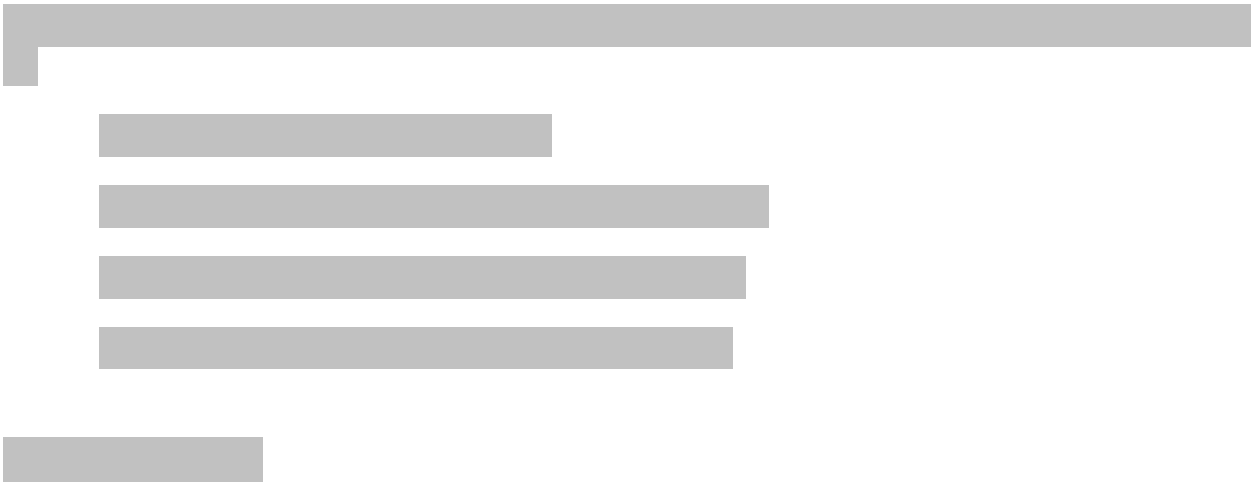
Test-form item #27
SCK.ISS.3

Test-form item #28
KCS.RPD.4

Test-form item #29
CCK.PO.2



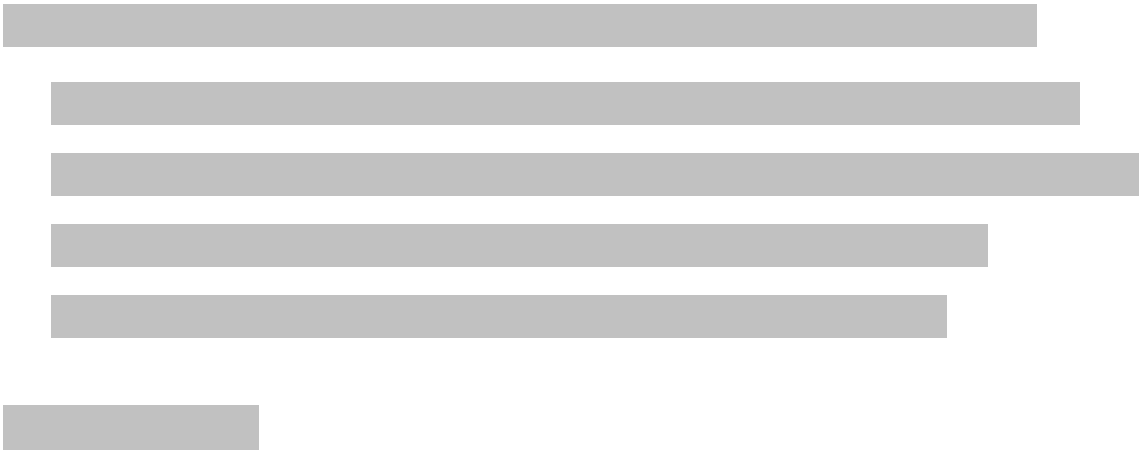
Test-form item #30
CCK.ES.5



Test-form item #31
SCK.ISS.6

Test-form item #32
SCK.ISS.5

Test-form item #33
CCK.PO.9



Test-form item #34
KCS.RPD.5

Test-form item #35
CCK.SMW.7

Appendix B. Administration Instructions at the Start of the Online Assessment

FLORIDA STATE UNIVERSITY

SURVEY OF PRIMARY GRADES MATHEMATICS KNOWLEDGE FOR TEACHING

INSTRUCTIONS FOR COMPLETING THIS QUESTIONNAIRE:

This questionnaire may take you as much as one hour to complete. You may find it useful to have pencil and scratch paper available as you complete this.

This is a questionnaire developed by mathematics educators to measure depth of knowledge for teaching primary grades mathematics. You may notice that these questions are designed to align with the way that a teacher needs to know mathematics in order to teach it.

In completing this questionnaire, you should not spend more than 1 or 2 minutes on any question. Imagine you are responding to real classroom situations, and select the answer that most closely matches what you would do, say, or answer at that moment.

Please answer questions based on your own knowledge. In other words, we request that you do not consult any external references (e.g., books, the Internet, or your colleagues) in order to respond to the questions. We are using the honor system to trust that you will comply with this request.

We recommend that you make every effort to complete the questionnaire in one session. If you are interrupted and must step away, your responses will be saved, and you may use the same survey link to finish at a later time. If you return to the survey from the same computer (and you have not cleared the computer's cookies in the interim), you will be able to resume your session, picking up where you left off; otherwise, the survey system will have you start from the beginning as a new session.

Once you have registered your response to each question and pressed the 'Next' button, you will *not* be able to revisit your responses to questions. In addition, please note that the back button (←) on your internet browser will *not* bring you to the previous question(s).

We are not using the information gathered by this questionnaire to evaluate any person individually. Your name will not be associated with any reporting of these data, and your responses will not be shared with your principal or your district.

Thank you for your participation. If you have any questions or concerns, please contact:

Rob Schoen, Principal Investigator
rschoen@lsi.fsu.edu

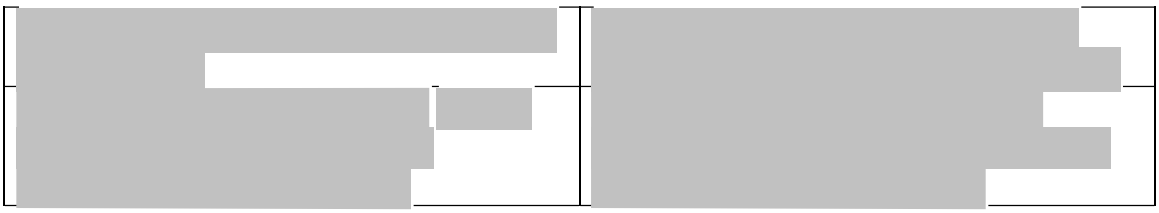
Amanda Tazaz, Project Director
atazaz@lsi.fsu.edu

Survey Completion
0% 100%

Next

Appendix C. Further Specification and Exemplars of Scoring Criteria for Constructed Response Items CCK.SMW.6 and SCK.ISS.2

CCK.SMW.6



iii.

SCK.ISS.2

was 45.

Appendix D. Proportion of Teacher Responses by Item

Table D.1. Proportion of Teacher Responses by Item (N = 387)

Item	Test-form item #	Final-scale item #	Correct response	Most frequent incorrect responses			
			Response (%)	Response (%)	Response (%)	Response (%)	Response (%)
KCT.RPD6	1	1*	B (.79)	D (.13)	A (.05)	C (.03)	
KCT.LG1	2	2*	A (.75)	D (.18)	B (.04)	E (.02)	C (.01)
CCK.EE1		3*	(.18)	NR (<.01)			
CCK.EE1a	3a		N (.66)	Y (.34)	NR (<.01)		
CCK.EE1b	3b		N (.70)	Y (.30)	NR (<.01)		
CCK.EE1c	3c		Y (.98)	N (.02)	NR (<.01)		
CCK.EE1d	3d		N (.78)	Y (.22)	NR (<.01)		
CCK.EE1e	3e		N (.42)	Y (.58)	NR (<.01)		
CCK.EE1f	3f		Y (.97)	N (.03)	NR (<.01)		
CCK.EE1g	3g		N (.67)	Y (.33)	NR (<.01)		
CCK.EE1h	3h		N (.67)	Y (.33)	NR (<.01)		
CCK.SMW8	4		D (.39)	B (.31)	C (.21)	A (.09)	NR (<.01)
CCK.SMW6	5	4*	(.21)	NR (.01)			
KCC.NPT14	6	5*	D (.37)	A (.56)	B (.05)	C (.01)	NR (.01)
CCK.EE2	7	6*	B (.57)	A (.18)	C (.13)	D (.13)	NR (.01)
KCC.NPT12	8	7*	A (.80)	D (.11)	B (.06)	C (.03)	NR (.01)
SCK.ISS2	9	8*	(.59)	NR (.01)			
CCK.ES3a	10a		8 (.52)	10 (.46)	9 (.02)	1 (<.01)	NR (.01)
CCK.ES3b	10b		12 (.96)	10 (.02)	11 (.02)	8 (.01)	NR (.01)
CCK.ES3c	10c		9 (.53)	10 (.45)	8 (.01)	NR (.01)	1 (<.01)
CCK.ES3d	10d		10 (.53)	11 (.44)	9 (.02)	1 (.01)	NR (.01)
CCK.ES3	10	9*	(.49)				
SCK.ISS1	11	10*	A (.41)	D (.30)	C (.15)	B (.14)	NR (.01)
KCC.NPT15	12	11*	C (.52)	D (.20)	B (.16)	E (.08)	A (.03)
KCT.LG2	13	12*	A (.61)	B (.23)	C (.08)	D (.08)	NR (.01)
CCK.PO7	14	13*	D (.56)	E (.20)	B (.13)	C (.09)	A (.03)
SCK.ISS4	15	14*	E (.54)	C (.16)	D (.16)	B (.12)	A (.02)
KCC.NPT1	16	15*	D (.56)	B (.30)	C (.07)	A (.04)	E (.02)
SCK.CMMI2	17	16*	D (.44)	A (.35)	B (.19)	C (.02)	NR (.01)
SCK.MSWP1	18	17*	C (.64)	B (.23)	D (.07)	A (.06)	NR (.01)
SCK.CMMI4	19	18*	B (.31)	A (.50)	E (.09)	D (.06)	C (.04)
CCK.PO8	20		C (.42)	A (.42)	B (.10)	D (.06)	NR (.01)
KCT.LG5	21	19*	B (.45)	C (.32)	D (.19)	A (.05)	NR (.01)
CCK.ES7	22	20*	C (.29)	B (.34)	E (.27)	D (.06)	A (.03)
SCK.MSWP3	23	21*	B (.67)	A (.19)	D (.11)	C (.03)	NR (.01)
SCK.CCMI3	24	22*	B (.56)	C (.39)	A (.03)	D (.02)	NR (.01)
CCK.ES2	25	23*	C (.43)	D (.33)	B (.21)	A (.03)	NR (.01)
SCK.MSWP2	26	24*	D (.69)	B (.13)	C (.10)	A (.08)	NR (.01)
SCK.ISS3	27	25*	D (.42)	C (.49)	A (.07)	B (.03)	NR (.01)
KCS.RPD4	28	26*	B (.65)	E (.19)	D (.13)	C (.03)	NR (.01)
CCK.PO2	29	27*	C (.70)	A (.14)	B (.11)	D (.03)	E (.03)
CCK.ES5	30	28*	B (.67)	C (.16)	D (.12)	A (.05)	NR (.01)
SCK.ISS6	31	29*	C (.55)	B (.20)	A (.16)	D (.08)	NR (.01)
SCK.ISS5	32	30*	D (.83)	C (.10)	B (.04)	A (.03)	NR (.01)
CCK.PO9	33	31*	C (.74)	A (.12)	B (.12)	D (.03)	NR (.01)
KCS.RPD5	34	32*	A (.63)	E (.28)	B (.04)	C (.03)	D (.02)
CCK.SMW7	35		D (.08)	A (.91)	C (.01)	NR (.01)	B (<.01)

Note. Proportions may not sum to 1 because of rounding; Test-form item # = the item index from the original test; Final-scale item # = the newly generated item number after item recoding (we added * after each final-scale item number).