DEVELOPING AN ASSESSMENT INSTRUMENT TO MEASURE EARLY ELEMENTARY TEACHERS' MATHEMATICAL KNOWLEDGE FOR TEACHING

ABSTRACT

This study reports on the development and field study of K-TEEM, a Web-based assessment instrument designed to measure mathematical knowledge for teaching (MKT) at the early elementary level. The development process involved alignment with early elementary curriculum standards, expert review of items and scoring criteria, cognitive interviews with practicing teachers, a field test involving 405 practicing teachers, and data modeling using a Rasch model. Several examples of MKT at the early elementary level are provided, and some of the challenges and decisions made during the process of item and scale development are discussed. Rasch model results indicate good model fit and adequate reliability, and the model accounts for more than 75% of the variance in the data. The K-TEEM assessment instrument may fill an important gap in the set of tools available to researchers for program evaluation and empirical investigation of teacher knowledge.

Robert C. Schoen
Wendy Bray
FLORIDA STATE UNIVERSITY

Christopher Wolfe
SAINT LEO UNIVERSITY

Amanda Tazaz
FLORIDA STATE UNIVERSITY

Lynne Nielsen
LOUISIANA TECH UNIVERSITY

THE ELEMENTARY SCHOOL JOURNAL
Volume 118, Number 1. Published online August 8, 2017
 0013-5984/2017/11801-0003$10.00

THERE is widespread agreement among scholars involved with research in teacher education that teachers' influence on their students' learning depends on teachers' subject-matter knowledge and their ability to draw on that knowledge in the practice of teaching (Borko & Putnam, 1996; Ma, 1999; Moats, 2009; Shulman, 1986). Speaking specifically of mathematics, Fennema and Franke (1992, p. 147) wrote, "Some scholars suggest that since one cannot teach what one does not know, teachers must have in-depth knowledge not only of the specific mathematics they teach, but also of the mathematics their students are to learn in the future." Consistent with the premise that teachers cannot teach what they do not know, the theory of change in most teacher professional development (PD) in mathematics and science posits that teacher PD has a direct effect on teacher knowledge, and this direct effect results, indirectly, in improvements to classroom instruction and increases in student learning (Smith & Banilower, 2006). Guided by this theory of change, many teacher PD programs have made it a primary goal to increase teachers' subject-matter knowledge (Garet, Heppen, Walters, Smith, & Yang, 2016; Sowder, 2007).

Confirmation of the link between teacher knowledge and student achievement in large-scale studies has had limited success, and the extant positive results seem disproportional to the firm beliefs and strong rhetoric in the broader literature on teacher education (Carlisle, Kelcey, Rowan, & Phelps, 2011; Hill, Ball, Blunk, Goffney, & Rowan, 2007; Hill, Rowan, & Ball, 2005; National Mathematics Advisory Panel, 2008; Rockoff, Jacob, Kane, & Staiger, 2011; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). There are several plausible explanations for the shortcomings in the evidence. Of course, one explanation could be that teachers' subject-matter knowledge is simply not as important a factor in teaching and learning as scholars believe it to be. Yet another explanation could be our limited ability to identify or measure the facets of teacher knowledge that matter.

Further development and refinement of assessment instruments designed to measure teacher knowledge may result in the creation of those critically important tools needed by researchers and evaluators to measure the effect of teacher PD programs on a large scale and gain insight into those facets of teacher knowledge associated with student learning. Development of reliable instruments that are valid for large-scale use in rigorously designed studies is difficult and resource intensive (Hill, Sleep, Lewis, & Ball, 2007). If a construct of teacher knowledge is ill-defined or poorly aligned with the type of knowledge that is most effective in supporting student learning, then a test designed to measure that construct may fail to detect the facets of teacher knowledge that matter for student learning. Furthermore, an assessment instrument can suffer from limitations due to construct irrelevance or construct underrepresentation (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), which could lead to failure to detect the association between teacher knowledge and student learning.

The purpose of this article is to describe and discuss a method used over the course of 1 year to develop an assessment instrument to measure teacher knowledge. Attempts were made to align the content of the assessment instrument with the mathematics early elementary teachers are expected to teach and the goals of two mathematics PD programs. We describe an iterative and overlapping process

of item writing and revision, expert review, use of test items in cognitive interviews with practicing teachers from the target population, and data modeling. The process was designed to continually strive toward clarification of the construct we were trying to measure and to minimize construct-irrelevant variance in the resulting data. Because the resulting assessment instrument is designed specifically to measure knowledge for teaching early elementary mathematics, we refer to this assessment instrument as *K-TEEM*.

Why Focus on Knowledge for Teaching Early Elementary Mathematics?

Following decades of research on general pedagogical knowledge needed for teaching, Shulman (1986, p. 9) introduced the construct of pedagogical content knowledge (PCK), which he described as "the particular form of content knowledge that embodies the aspects of content most germane to its teachability." Elaborating on Shulman's theory and applying it within mathematics, Ball, Thames, and Phelps (2008) theorized a delineation of multiple facets within the domains of content knowledge and PCK in a construct they named "mathematical knowledge for teaching" (MKT).

At present, the most well-known and widely used measures of MKT are those derived from the item bank developed through the Study of Instructional Improvement (SII) and Learning Mathematics for Teaching (LMT) projects (Hill, Schilling, & Ball, 2004; LMT, 2004). Arguably, the second most widely known instrument used by program evaluators and researchers to measure MKT are the Diagnostic Teacher Assessment of Mathematics and Science scales (Bush, Ronau, Brown, & Myers, 2006; Saderholm, Ronau, Brown, & Collins, 2010). Campbell et al. (2014) recently developed another instrument designed to measure teachers' MKT—both content knowledge and PCK.

Although the number of high-quality items and scales that can be used efficiently to measure teachers' knowledge for teaching mathematics is on the rise, the content coverage in those existing scales tends to be most relevant to the content expected to be taught by upper elementary and middle-grades teachers. By design, the measures developed by Campbell et al. (2014) are aligned with the content in the standards for upper elementary and middle-grades mathematics. In a validity study of the SII/LMT items and scales, the LMT developers identified the early elementary subject matter (i.e., K–2) as an area with a need for further development (Hill & Ball, 2004; Seidel & Hill, 2003).

In our own initial efforts to measure the knowledge of teachers involved in a mathematics PD project, we administered a pretest in summer 2013 consisting of items gathered from the LMT item bank.[1] We searched the LMT item bank to select items that met the following criteria: (*a*) the content of the item focuses on the topic of number, operations, or algebraic thinking; (*b*) the numbers presented in the item involve whole numbers and do not involve common fractions, decimal fractions, or negative integers; and (*c*) items involve specific numbers and do not require teachers to interpret letters or other symbols as variables. Our search yielded 23 items that met these criteria, and all 23 items were used to construct a paper-and-pencil assessment.

Intending to use the scale to measure the effects of a PD program, we administered the 23-item scale to more than 200 public school elementary teachers and math coaches in a single southeastern state as a pretest in summer 2013. Using this sample, the Cronbach's alpha reliability estimate for the 23-item scale was .61. The low reliability estimate and the limited number of items fitting our content specifications were the impetus for the development of an instrument designed to be a reliable measure of MKT that would be valid for use with the general population of U.S. teachers working with early elementary-grades students.

Theoretical Framework

Hill et al. (2005, p. 373) define MKT as "the mathematical knowledge used to carry out the *work of teaching mathematics*" (emphasis in original). We conceptualize the work of teaching mathematics to include interactions with students in the classroom setting as well as activity in related contexts, such as planning for teaching and reflecting on teaching and learning (Ball et al., 2008; Goldsmith, Doerr, & Lewis, 2014). Ball et al. (2008) suggest several subdomains that compose the larger domain of MKT, including common content knowledge (CCK), specialized content knowledge (SCK), knowledge of content and students (KCS), knowledge of content and teaching (KCT), horizon content knowledge (HCK), and knowledge of content and curriculum (KCC).

In designing the K-TEEM scale, we used an iterative process (*a*) to identify important and measurable facets of knowledge for teaching early elementary mathematics, (*b*) to sort these various facets of knowledge into the theoretical categories of the existing MKT framework, and (*c*) to write items to try to yield insight into whether teachers have these facets of knowledge. The resulting K-TEEM instrument includes items that reflect four of the theoretical subdomains of MKT, CCK, SCK, KCS, and KCT. HCK involves knowledge of the mathematics topics that students will encounter in the future. For early elementary, one of those topics on the horizon might be rational numbers (e.g., fractions). We ruled out measuring HCK because our intent was to focus specifically on the topics students are expected to learn (and teachers are expected to teach). Because teachers in different places use different textbooks, and textbooks are continually revised, we chose to refrain from focusing on KCC. But we did use the Common Core State Standards for Mathematics (CCSS-M; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) as a general guideline for delineating content and determining what kind of topics are fair to expect most U.S. teachers to know. In the following sections, we briefly describe our working definitions of each of these four subdomains of MKT and provide original examples of types of knowledge we attempted to measure within each subdomain.

Common Content Knowledge

Ball et al. (2008, p. 399) define CCK as "the mathematical knowledge and skill used in settings other than teaching." For example, most people who work with

mathematics regularly have mathematical knowledge that allows them to solve equations such as $200 - x = 186$ in a variety of ways. This knowledge is likely to be useful in the act of teaching, but it is not uniquely useful to the work of teaching mathematics.

Our working definition of CCK also includes knowledge of the formal use of mathematical vocabulary and conventions of notation commonly acknowledged in the broader mathematics community. For instance, this knowledge might involve an awareness of a distinction in meaning between the words *expression* and *equation* in mathematics. As another example, a person with strong CCK might be expected to recognize the commutative property of addition by name or understand why equation (1), intended to explain how a person might add 34 and 16, violates generally acknowledged conventions of formal mathematical notation:

$$34 + 6 = 40 + 10 = 50. \quad (1)$$

Specialized Content Knowledge

Ball et al. (2008, p. 400) define SCK as "the mathematical knowledge and skill unique to teaching." The authors discuss SCK as a way of knowing about mathematics that is uniquely useful in teaching and not necessarily required or useful by persons working in other professions that use mathematics. They offer an example of how teachers use decompressed knowledge of the mathematics they teach to efficiently size up the conceptual basis of a student's error.

Consistent with our view of teaching as including planning for instruction and participating in professional discussions with other teachers, our working conceptualization of SCK includes the knowledge of shared, professional vernacular related to the teaching and learning of mathematics. For example, we believe that early elementary-grades teachers with strong SCK are aware of the differences in semantic structure among addition and subtraction word problems, various equations that would model the structure of the problem, and terms to describe these differences (Carpenter, Fennema, Franke, Levi, & Empson, 1999; Fuson, 1992; Nesher, Greeno, & Riley, 1982; Verschaffel, Greer, & De Corte, 2007). The following problem is considered a compare-type problem with the difference unknown; it would not be considered to be a "change-unknown" problem: "Luca has six trophies. Sofia has four trophies. How many more trophies does Luca have than Sofia?" The Operations and Algebraic Thinking domain in the CCSS-M references taxonomies based on these factors in both first and second grades (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Thus, understanding the CCSS-M requires a teacher to be aware of these distinctions.

Whereas this knowledge of professional vernacular may be important in teaching children mathematics, professionals who do not teach children mathematics (or study how people learn mathematics) are unlikely to know the vernacular or find it useful to know this in their own professional work. This is analogous to reading teachers knowing specialized terms, such as *morphemes* and *phonemes*—terms that laypersons do not need to know to be able to read sufficiently well. It is analogous to medical doctors using Greek or Latin words to describe parts of

the body. The layperson need not know the vernacular used by doctors to identify body parts, but the professional vernacular enables efficient and precise conversations among professionals within the medical community.

Knowledge of Content and Students

KCS is "knowledge that combines knowing about students and knowing about mathematics" (Ball et el., 2008, p. 401). KCS is the amalgamated knowledge about how students think about mathematics that makes it possible for teachers to accurately predict or diagnose how students think about and interact with mathematics content. For example, Ball et al. (2008) include teachers' abilities to predict and categorize common errors made by learners as examples of KCS. Notice how the ability to predict that a given group of second-grade students will make a particular error (a matter of KCS) is categorically different from being able to recognize that an error has been made (a matter of CCK).

Teachers with high levels of KCS are able to anticipate the most common ways that learners with different levels of understanding will approach problems, and these teachers know which of the problems will generally be the easiest and most difficult for students to solve. For example, when presented with the equation $10 = 7 + 3$ and asked whether the equation is true or false, first- and second-grade students typically answer *false* (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016). Teachers with strong KCS (*a*) will know that students are likely to answer this question incorrectly and (*b*) will be able to explain why a student might think the equation is false.

As another example, consider the following word problem: "Iris had nine flowers. She gave some flowers to her mother. Now, she has three flowers. How many flowers did Iris give to her mother?" A teacher with strong KCS would expect many first- or second-grade students to write an equation structured like this: $9 - x = 3$. Teachers with low KCS for early elementary mathematics are often surprised to see young children use the $9 - x = 3$ equation structure to model this problem rather than using the way that most adults would think of it, which is to think in terms of the equation $9 - 3 = x$ (T. Carpenter, personal communication, October 2, 2014).

Knowledge of Content and Teaching

Ball et al. (2008, p. 401) discuss KCT as a type of knowledge that "combines knowing about teaching and knowing about mathematics." KCT is knowledge that facilitates skillful instructional design—the design and sequencing of specific mathematics problems and experiences to provoke particular aspects of student thinking and accomplish specific instructional goals. Instructional decisions that draw on KCT require "coordination between the mathematics at stake and the instructional options and the purposes at play" (Ball et al., 2008, p. 401).

One fundamentally important idea in mathematics that early elementary students are expected to learn is the notion of place value (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). A typical related task in textbooks involves presenting students with a numeral, such as 50, and directing the student to circle the numeral in the tens place. If the student

circles the 5, the teacher or test developer infers that the student has some understanding of place value. If not, the teacher or test developer infers that the student does not understand place value.

Consider the following problem: "There are 5 people at Sally's birthday party. Each person eats 10 pieces of candy. How many pieces of candy are eaten?" This problem can be considered to be a multiplication problem, but there is another important aspect of this problem related to place value in a base 10 number system. Teachers with high levels of KCT can recognize the grouping-by-10s structure in the word problem, and they can see how this problem would be a useful tool for the teaching and formative assessment of place-value understanding. We have observed in our work that not all teachers or school administrators notice the relation between the situation in this word problem and opportunities for students to learn about or demonstrate their understanding of place value.

Description of the K-TEEM Test Development Process

We developed items intended to measure teachers' MKT in a way that focused on the types of knowledge that teachers in the early elementary grades may need to know to teach number, operations, and algebraic thinking. Table 1 presents the major phases of the development process we used.

Item Generation

To define the content focus for the assessment instrument, we started with a close review of the CCSS-M (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and the learning goals for teachers in two PD programs: Cognitively Guided Instruction (Carpenter et al., 1999; Fennema, Carpenter, Levi, Franke, & Empson, 1999), and Thinking Mathematics (Bodenhausen, Denhart, Gill, Kaduce, & Miller, 2014). These two PD programs both focus on number and operations—the mainstay of the elementary mathematics curriculum. Both programs encourage teachers to use the following strategies to guide their instructional decisions: Use story problems to introduce mathematical concepts, build on students' existing and intuitive understanding of mathematical ideas, emphasize both conceptual and procedural learning, and make continual adjustments to the instructional plan based on ongoing formative assessment.[2] Both programs are aligned with student learning expectations identified in the CCSS-M.

The development team used the CCSS-M as a touchstone to provide guidelines for avoiding overalignment of the instrument to the specific PD programs being evaluated (Slavin & Madden, 2011). We targeted the content found at the intersection of the learning goals of the teacher PD programs and the learning goals for students described in the early elementary CCSS-M.

After identifying and naming specific learning goals of the PD programs that might be both measurable and consistent with the CCSS-M, we considered how these facets of knowledge might map onto the theoretical framework for MKT proposed by Ball et al. (2008) and Hill et al. (2005). We then established a target blueprint for the MKT instrument and drafted items in accordance with this blueprint.

Table 1. Major Phases of Instrument Development

Phase	Duration	Activities
Review existing instruments	8 months	Search for existing instruments aligned with the focus of the intervention being tested through review of extant literature and discussions with experts in field of mathematics teacher education and program evaluation
		Identify and select extant and available assessment items aligned with the focus of the intervention
		Field-test the available instruments with a sample from the target population
		Analyze resulting field-test data
Item generation	4 months	Review the aspects of MKT relevant to the teacher PD program and the CCSS-M
		Develop a target blueprint detailing types of items and number of items of each type
		Draft items in accordance with the target blueprint and item specifications
Item refinement	4 months	Review items by experts in mathematics, mathematics education research (including teacher education and student thinking), and practicing teachers
		Conduct cognitive interviews with practicing teachers
		Discuss notes and observations generated through cognitive interviews with the development team
		Revise or write new items based on cognitive interview findings
		Determine the final set of items to be included in field test
Field test	4 months	Transfer paper-based items to Web-based system
		Pilot-test Web-based system
		Full-scale field-test Web-based system
Scale refinement	3 months	Adjudicate responses for fill-in-the-blank items and scoring of responses for short answer items
		Develop and analyze Rasch-based models
		Determine final set of items based upon model results

Note.—Some of the activities in the test and item development process occurred in an iterative and overlapping fashion. The development period lasted a total of approximately 15 months from start to finish.

Items available to us through various existing instruments designed to measure facets of MKT were reviewed for inspiration (LMT, 2004; Rittle-Johnson, Matthews, Taylor, & McEldoon, 2011; Saderholm et al., 2010; Wheeler, 2010). Items from these sources were used with permission from their original authors and modified and adapted for use in this new instrument. The K-TEEM scale included one item that was adapted from each of the four referenced sources.

Item types for the K-TEEM included multiple-choice, fill-in-the-blank, and constructed-response items. We avoided creating multiple items referencing the same prompt (e.g., item sets, testlets) as an attempt to maintain the independence of each item in the test. Multiple-choice items were designed to include one and only one correct response. Both of those decisions were made in support of the goal to simplify scoring and interpretation. The use of *all of the above, undecided*, and *none of the above* options in multiple-choice items was discouraged. We expended considerable effort to write only response options that the practicing teachers would consider plausible or that otherwise reflected their thinking (Haladyna, Downing, & Rodriguez, 2002). The item-generation phase occurred over a period of 4 months of daily effort with a team of four item writers.

Item Refinement

After the initial drafting of items, we used two activities iteratively to vet and refine the item bank: (*a*) consultation and discussion with experts and (*b*) cognitive interviews with early elementary teachers. Both activities will be described briefly in the following sections.

Expert review. Experienced classroom teachers, teacher PD leaders, and other experts in mathematics and mathematics education reviewed the draft items and provided feedback. We specifically elicited feedback on (*a*) what the experts thought each item was measuring, (*b*) potential issues related to each item's clarity and validity, (*c*) what to accept as a correct answer for the item, and (*d*) how difficult the items would be for respondents. Major goals at this stage were to identify whether the questions were well posed and to make sure that the determination of correct answers would be acknowledged by all experts in the field, regardless of potential differences of opinion (Downing, 2006).

We decided whether to keep, eliminate, or revise items based on this initial round of expert feedback. From the bank of approximately 70 items developed through this process, 55 items were judged to have valid correct answers and to be aligned with the content of the draft test blueprint. These items were then advanced to the next phase of development to be used in the cognitive interviews with a small set of early elementary teachers. Organized by MKT subdomain and further subcategories within the subdomains, the number of draft items available for the cognitive interviews are presented in Table 2.

Cognitive interviews. Cognitive interviews involve asking respondents to perform tasks in the presence of an interviewer and to verbalize their thought processes during and after they perform the tasks (Desimone & Le Floch, 2004). The

Table 2. Number of Draft Items by MKT Subdomain and Subcategory at the Start of Cognitive Interviews

Subdomain and Subcategory	Number of Items
Common content knowledge:	
Meaning of the equal sign and related notation	7
Properties of operations	7
Solve problems in many ways	4
Specialized content knowledge:	
Evaluating the validity or generalizability of student strategies[a]	4
Naming student strategies	6
Naming word problem types	5
Writing word problems[b]	3
Knowledge of content and students:	
Predicting student strategies	4
Relative problem difficulty	5
Matching strategies and problems[a]	5
Knowledge of content and teaching:	
Selecting word problems in service of specific instructional goals	5
Total	55

[a] These categories were dropped or reconceptualized based on the information gathered in the cognitive interviews.

[b] This category was dropped due to concerns about whether respondents would use their own knowledge or consult external references in the Web-based, self-paced format.

interviewer observes the respondent and asks questions to further clarify how the respondent was thinking about the items and responses.

Using data collected through cognitive interviews, we used a critical eye to gain insight into whether the items were consistently yielding information about the types of knowledge the items were intended to measure. As interviewees revealed reasons for their answers, we learned about the aspects of the items that respondents tended to overlook, how they interpreted the questions, and how they responded.

Three of the authors of this article (who were also developers of the items on the K-TEEM test) served as the interviewers for the cognitive interviews. All three had intimate knowledge of what each item was designed to measure as well as prior experience with using questioning techniques to probe the details of teachers' thinking about mathematics and mathematics teaching and learning.

In the first round of cognitive interviews, five teachers participated in interviews lasting between 90 and 120 minutes. In preparation for the interviews, we set a maximum time limit of 120 minutes to be considerate of participants' time. Interviewers were instructed to terminate the interview earlier if the interviewer perceived the interviewee to be experiencing significant fatigue or frustration. Interviewers noticed that teachers tended to show signs of fatigue in the cognitive interviews at 75 to 90 minutes. These signs included sighs, comments about being tired, and flipping through the booklet of questions to see how many questions remained. Sometimes the interviewees said directly that they were tired and ready to stop. Along with field notes taken by the interviewers, each interview was audiotaped for subsequent review and analysis by the item development team.

After the first round of cognitive interviews, the data generated from the interviews were compiled and analyzed. Item by item, the development team carefully examined and compared how interviewees responded. Based on these detailed analyses, we gained insight into aspects of items that made them easy, difficult, confusing, time-consuming, enjoyable, frustrating, or cognitively demanding. Questions about properties of operations, for instance, typically invoked feelings of frustration, whereas the videos of children solving problems seemed to have an invigorating effect on the teachers. The interviews also informed the selection and editing of response options for individual items, and we used all of this information to revise, eliminate, and create new items.

Using the revised items, we conducted a second round of cognitive interviews with six additional elementary teachers. The second round of interviews followed the same process as the first, including the sharing of audio recordings and extensive follow-up conversations among the development team (i.e., the authors of this article). Following the cognitive interviews, we focused considerable attention on confirming plausible incorrect responses that reflected the thinking observed among the target population and limiting the number of response options in multiple-choice items accordingly. We sought to limit the amount of time required for teachers to read the items. We translated several vocabulary terms in the draft items to synonyms that were used by the teachers, and we edited multiple-choice response options to have similar grammatical structure, vocabulary, and length (Haladyna et al., 2002). Above all, items were edited and proofed endlessly. The item-refinement phase occurred over a period of 4 months of intensive effort and critical feedback and discussion.

Field Test

After the item-refinement phase was complete, the bank of remaining items consisted of items that had not been eliminated on the grounds that they were too easy, too difficult, or too time-consuming, or they failed to illuminate whether teachers had the type of knowledge or ability we sought to measure. Partly influenced by the observations in the cognitive interviews that teachers showed clear signs of fatigue after 75 to 90 minutes, we aimed to set the number of items such that teachers were likely to complete the test in 60 minutes or less. Thus, we set a target to keep approximately 35 to 40 items. To decide which items to keep, we examined how many items remained in each of the test blueprint categories. For the categories with more than three items in them, we identified pairs or groups of items with very similar structure that were designed to measure the same facet of knowledge. We identified and retained the items from these pairs or groups that seemed to be the most effective and efficient at illuminating the targeted types of knowledge with teachers in the cognitive interviews, and we removed the others. This process yielded 40 items.

These 40 items were used to create an online version of the instrument using the Qualtrics software, a Web-based platform that afforded a multimedia approach. Some items include image files depicting student work or videos of students solving mathematics problems. The test blueprint in Table 3 shows the revised categories of items represented on the K-TEEM and the final number of items in each category after the spring 2014 Web-based field test and data analysis. Five of the 40 items were removed after data collection, and the reasons are discussed in the following section. These steps in the field-test phase occurred over a period of 6 months, not including the time required to recruit participants.

Table 3. Blueprint of Items by MKT Subdomain and Subcategory Included in Final Analyses

Subcategory of Items by Subdomain of MKT (Item Code)	Number of Items
Common content knowledge:	
Meaning of the equal sign and related notation (ES)	5
Properties of operations (PO)	4
Solve problems in many ways (SMW)	2
Specialized content knowledge:	
Interpreting student strategies (ISS)	4
Naming student strategies (NS)	4
Naming word problem types (NPT)	5
Knowledge of content and students:	
Predicting student strategies (PS)	3
Relative problem difficulty (RPD)	4
Knowledge of content and teaching:	
Selecting word problems in service of specific instructional goals (LG)	4
Total	35

Note.—There were a total of 40 items on the questionnaire. After the field-test data were collected, five items were dropped in the process of data analysis. Dropped items were placed in the following subcategories: KCS-Predicting student strategies, CCK-Solve problems in many ways, SCK-Naming problem types, and SCK-Interpreting student strategies. The final version of the questionnaire used in the analytic sample has 35 items.

Scale Refinement

The 40-items used in the spring 2014 field test consisted of 30 multiple-choice items, three fill-in-the-blank items, and seven constructed-response items. The multiple-choice items were scored in accordance with an a priori determination of correct responses. After the field-test data were collected, the development team worked as an adjudication committee to examine all of the responses to the fill-in-the-blank answers to determine the set of possible correct answers to those questions.

We created rubrics to score each of the constructed-response items. These rubrics were drafted based on the responses observed in the cognitive interviews and then refined through an iterative process of scoring, comparing scores, and refining the scoring criteria. After the first draft of the rubrics was created, the members of the development team scored a subset of items individually. These scores were then compared, and all discrepancies were discussed and resolved by the full group. Two of the seven constructed-response items were dropped from the scale during the scoring process due to a combination of difficulty in defining objective scoring criteria and shortcomings in achieving a sufficiently high percentage of exact agreement in rating. Full consensus on every score was achieved in every case for the remaining items.

All items on the K-TEEM test were ultimately scored dichotomously (i.e., correct, incorrect), and statistics for the remaining 38 individual items were generated using Rasch (1960/1980) models. The Rasch model output data identified three items with poor model fit (see the Results section for further discussion). Those three items were subsequently removed. The final 2014 K-TEEM scale includes 35 items involving a mix of multiple-choice (27 items), fill-in-the-blank (2 items), and constructed-response (6 items) formats. Scoring, data modeling, and interpretation of results occurred over a period of 4 months.

Validation Framework

Kane (2006) provides a useful way of framing test validation in terms of two basic components: the interpretation argument and the validity argument. The interpretation argument is focused on what a test score means or, put another way, what can we infer that a score tells us about the test taker. The validity argument focuses on how a test is used and whether the use and inference thereof is appropriate and defensible. With respect to the process of test development and validation, Kane further distinguishes among the development stage and the appraisal stage. Our current work is situated in the development stage. As such, we focus most of our attention on the interpretation argument while attempting to provide clear direction for the subsequent appraisal and validity argument.

Our current work focuses on building an argument to support the interpretation of the test score, and subsequent work will appraise the ability of the K-TEEM to serve its intended use. In the previous sections, we defined the domain of interest (i.e., MKT at the early elementary level in the domain of number and operations, and equality), offered a test blueprint and other test specifications for the K-TEEM, and described the development process, including an iterative process of subjecting items to expert review and cognitive interviews. In the following sections, we will describe a feasibility test and share related findings.

Description of the Sample and Setting

The sample for this study includes early elementary-grades teachers of mathematics (kindergarten through second grade) and instructional support personnel (e.g., math coaches, intervention specialists) who signed up to take part in a teacher PD program in mathematics. All teachers in the sample worked as teachers in the state of Florida. The data for this study were gathered in spring 2014, during the last 9 weeks of the school year. The Web-based questionnaire was administered with participants involved in two separate randomized controlled trials of mathematics PD programs serving teachers in early elementary grade levels ($n = 405$). All teachers in both the cognitive interviews and the field-test phases were remunerated for their participation.

Sample 1. Approximately half of the teachers (i.e., Sample 1) completed the online instrument while signing up to participate in a randomized controlled trial that would evaluate the impact of a 10-day summer workshop based on the Thinking Mathematics program (Bodenhausen et al., 2014). Sample 1 data used in this study were collected prior to random assignment and delivery of PD, so the teachers were not aware of what condition they would be assigned (i.e., treatment, control), and they had not yet participated in any PD offered through the program. The teachers in Sample 1 ($n = 206$) represented 26 school districts, spanning the full geographic range of the state and including urban, suburban, and rural areas. Eligibility for enrollment in Sample 1 was constrained to those school districts that met the criterion for being high needs, as defined by student enrollment at or above the level of 50% of students qualifying for free or reduced-price lunch.

The average number of years of experience among the teachers in Sample 1 was 10.75 ($SD = 7.75$) years. The minimum number of years of experience was 1, and the maximum number of years of experience was 33. Three of the 206 Sample 1 teachers (1.5%) reported having earned a college degree specifically in mathematics or mathematics education. Sample 1 teachers predominantly identified as female (93.7%). Sample 1 consisted mostly of classroom teachers (96%), with only eight participants (4%) identifying with an instructional support role.

Sample 2. The remaining teachers (i.e., Sample 2) in the 2014 field study were completing an end-of-year posttest for the first year of a 2-year-long randomized controlled trial evaluating a PD program based on Cognitively Guided Instruction (Carpenter et al., 1999). Approximately half of the Sample 2 teachers were in the treatment condition, and the other half were in a practice-as-usual control condition. Sample 2 teachers were from two school districts in the same state as the Sample 1 teachers. One of those school districts was a very large district with urban, suburban, and rural areas within it. The other district was a medium-sized school district serving primarily suburban areas. Some teachers in Sample 1 were from the same two districts as the teachers from Sample 2, but none of the individual teachers were included in both samples.

The average number of years of experience among the teachers in Sample 2 was 11.63 ($SD = 8.84$) years. The minimum number of years of experience was 1, and the maximum number of years of experience was 48. Three of the 199 Sample 2 teachers (1.5%) reported having earned a college degree specifically in mathematics or mathematics education. Sample 2 teachers predominantly identified as female

(98.5%). Sample 2 consisted mostly of classroom teachers (91%), with 18 participants (9%) identifying with an instructional support role.

Analytic Strategy

We use the Rasch (1960/1980) model to obtain both the item-difficulty and the person-ability estimates. The joint maximum likelihood estimation of both the person-ability and item-difficulty parameters in the Rasch model provides a distinct advantage over simply assigning a person's ability as the percentage of items answered correctly. Using differences in the difficulty levels for individual items, the Rasch model is able to differentiate between similar respondent patterns occurring at separate points in the scale. By having different spacing between Rasch-based scores, we better reflect the true ability differences between people. All analyses for this study were conducted with the Winsteps computer program (Linacre, 2016).

Allowing us to place item difficulty and teacher knowledge on the same scale, the Rasch approach is both convenient and easy to interpret. It also provides some improvement in precision, as precision is maximized in the center of the Rasch score distribution versus in the tails for raw score scales (Bond & Fox, 2007). We also chose the Rasch approach because it helps us to evaluate whether the underlying construct of teacher mathematics knowledge is sufficiently unidimensional. Use of a raw score would make an implicit assumption of unidimensionality, whereas Rasch statistics allow us to evaluate the tenability of that assumption. This is an important advantage over using raw scores, given that a major contribution of the MKT is the broad range of knowledge for the instruction of mathematics.

Field-Test Findings

The Web-based field test involved 405 practicing teachers who completed the K-TEEM in spring 2014. Most participants required between 30 and 50 minutes to complete the test. There were a few technical problems, mostly involving participants experiencing difficulties logging into the system or accessing videos embedded in the items. These technical problems were resolved in all known cases.

Item-Level Performance across Samples

The respondents in the two samples differed across multiple items in the observed probability of a correct response. Table 4 displays descriptive statistics for the overall sample and two subsamples. Mean item-level scores in the final set of 35 items for Sample 1 ranged from 12% (Item 25) to 85% (Item 16) of the teachers correctly responding to an item. Less than 20% of Sample 1 ($n = 206$) answered correctly on Items 14, 25, and 35. In contrast, more than 70% of this sample answered correctly on Items 1, 16, 20, and 22. The item-level percentage-correct responses for Sample 2 ($n = 199$) ranged from 18% (Item 14) to 85% (Item 22). Less than 20% of Sample 2 answered correctly on Item 14. More than 70% of Sample 2 teachers answered correctly on Items 1, 4, 9, 12, 16, 20, 22, and 37.

Table 4. Average Correctness and Individual Correlation within Each Sample

		Sample 1 (n = 206)			Sample 2 (n = 199)		
Item	Item Code	Mean	*SD*	PB	Mean	*SD*	PB
1	KCSRPD6	.73	.446	.22	.81	.390	.25
2	KCSPS2	.37	.485	.23	.47	.500	.31
3	CCKES3	.24	.430	.30	.37	.483	.45
4	KCTLG1	.65	.477	.31	.79	.409	.32
5	SCKNPT1	.45	.499	.32	.57	.497	.50
6	KCSRPD1	.63	.484	–	.56	.497	–
7	KCTLG2	.59	.492	.34	.69	.462	.31
8	CCKPO7	.45	.499	.38	.38	.487	.36
9	KCSRPD5	.68	.469	.32	.77	.423	.37
10	CCKES2	.36	.481	.22	.31	.464	.28
11	KCSPS5	.43	.497	.23	.52	.501	.32
12	SCKNPT12	.65	.477	.23	.77	.423	.41
13	KCSPS6	.67	.470	.19	.61	.488	.29
14	CCKES7	.18	.383	.24	.18	.382	.32
15	SCKNS3	.29	.455	.14	.38	.487	.33
16	SCKSMW6	.85	.362	.34	.76	.429	.34
17	CCKSMW6	.23	.421	.47	.32	.466	.55
18	KCSRPD3	.49	.501	.17	.57	.496	.20
19	SCKISS3	.25	.435	.22	.38	.486	.42
20	CCKES10	.74	.438	.35	.82	.386	.27
21	CCKPO3	.21	.412	–	.24	.426	–
22	SCKNPT13	.77	.424	.17	.85	.359	.31
23	SCKISS4	.45	.499	.30	.49	.501	.45
24	CCKPO2	.67	.472	.39	.66	.475	.31
25	CCKSMW5	.12	.322	.34	.25	.432	.61
26	KCTLG3	.30	.459	.24	.39	.489	.31
27	SCKNPT14	.43	.496	.36	.52	.501	.35
28	SCKNS2	.44	.498	.17	.44	.498	.28
29	SCKNS6	.41	.493	.39	.49	.501	.35
30	CCKES5	.57	.496	.29	.66	.474	.27
31	CCKPO6	.40	.491	.34	.42	.495	.23
32	KCSRPD4	.39	.488	.34	.58	.494	.41
33	SCKISS1	.07	.256	–	.11	.308	–
34	CCKPO5	.36	.482	.28	.35	.477	.27
35	KCTLG4	.18	.387	.22	.22	.416	.41
36	SCKISS2	.50	.501	.38	.59	.493	.44
37	SCKISS5	.60	.491	.34	.77	.419	.40
38	SCKNS7	.39	.488	.34	.45	.499	.41

Note.—Mean = average of percentage correct response within sample for each item; *SD* = standard deviation; PB = individual item correlation to the overall measure. The item coding scheme involves three fields. The first three letters are the code for the subdomain of MKT (e.g., CCK = common content knowledge). The next two or three letters correspond to the subcategory of knowledge in that domain (e.g., RPD = relative problem difficulty). The numeral at the end simply indexes the items in the item bank in that subdomain.

Rasch Model Fit

To determine whether the K-TEEM items fit the Rasch model, item fit mean square (MNSQ) was examined within and across each sample. Items were considered misfit if the MNSQ estimates were either less than 0.6 or greater than 1.4 (Bond & Fox, 2007). Low values of MNSQ may indicate redundancy with other items, whereas high values may indicate items out of sync with other items in the measure.

Table 5 displays the infit and outfit MNSQ values as well as item difficulties and discrimination parameters within each group. Across both samples, Items 6 and 21 demonstrated the worst fit to the model as well as the lowest correlation to the underlying construct. These two items were dropped from further analysis. Item 33 represented the hardest item in the test ($\Theta = 2.47$) and was eliminated for the negative impact that misalignment of item difficulty can have on individual person-proficiency estimates, to help eliminate items with overt guessing and to hone the dimensionality of the overall measure (Andrich & Marais, 2014). The remaining K-TEEM items had acceptable infit statistics within the .60 to 1.4 range for both

Table 5. Item Difficulty and Fit Statistics Within and Across Both Samples

		Sample 1 ($n = 206$)				Sample 2 ($n = 199$)				Overall ($n = 405$)			
Item	Item Code	Item	Infit	Outfit	Dis	Item	Infit	Outfit	Dis	Item	Infit	Outfit	Dis
1	KCSRPD6	−1.26	1.04	1.06	.92	−1.51	1.03	1.07	.97	−1.37	1.03	1.05	.96
2	KCSPS2	.39	1.05	1.05	.83	.32	1.07	1.07	.75	.36	1.06	1.05	.80
3	CCKES3	1.07	.98	.97	1.03	.81	.94	.89	1.18	.93	.95	.92	1.11
4	KCTLG1	−.88	.99	.99	1.03	−1.34	.97	.94	1.04	−1.07	.97	.95	1.06
5	SCKNPT1	.05	.99	.99	1.05	−.15	.87	.83	1.48	−.04	.93	.91	1.33
7	KCTLG2	−.59	.97	.99	1.11	−.78	1.04	.99	.94	−.67	1.00	.99	1.02
8	CCKPO7	.05	.94	.95	1.27	.73	1.03	.99	.95	.37	1.00	1.00	.99
9	KCSRPD5	−.99	.98	.97	1.05	−1.21	.95	.86	1.10	−1.09	.96	.91	1.09
10	CCKES2	.46	1.05	1.06	.83	1.1	1.08	1.12	.82	.76	1.09	1.11	.77
11	KCSPS5	.12	1.05	1.06	.74	.08	1.06	1.05	.78	.10	1.05	1.05	.77
12	SCKNPT12	−.88	1.05	1.04	.86	−1.21	.91	.84	1.16	−1.02	.98	.94	1.06
13	KCSPS6	−.97	1.07	1.09	.81	−.37	1.06	1.12	.79	−.69	1.08	1.14	.74
14	CCKES7	1.50	.99	1.03	1.00	1.97	1.01	1.06	.98	1.72	1.01	1.07	.97
15	SCKNS3	.81	1.09	1.18	.79	.73	1.05	1.08	.85	.77	1.06	1.12	.82
16	SCKSMW6	−2.03	.93	.84	1.08	−1.15	.96	1.07	1.05	−1.59	.96	1.04	1.05
17	CCKSMW6	1.16	.87	.78	1.22	1.07	.83	.77	1.36	1.11	.85	.77	1.29
18	KCSRPD3	−.11	1.09	1.11	.47	−.17	1.16	1.16	.43	−.14	1.12	1.13	.45
19	SCKISS3	1.02	1.04	1.07	.93	.76	.96	.97	1.10	.88	.99	1.01	1.02
20	CCKES10	−1.34	.94	.93	1.11	−1.55	1.00	.99	.99	−1.43	.97	.95	1.05
22	SCKNPT13	−1.48	1.06	1.11	.89	−1.79	.96	.89	1.05	−1.60	1.01	1.00	.99
23	SCKISS4	.05	1.00	1.01	.99	.23	.93	.92	1.25	.13	.97	.97	1.12
24	CCKPO2	−.95	.94	.91	1.19	−.59	1.01	1.16	.90	−.78	.97	1.06	1.03
25	CCKSMW5	2.02	.93	.76	1.08	1.47	.77	.65	1.35	1.70	.83	.68	1.20
26	KCTLG3	.76	1.02	1.06	.93	.68	1.05	1.13	.81	.72	1.03	1.09	.89
27	SCKNPT14	.16	.96	.95	1.18	.08	1.03	.99	.91	.12	.99	.97	1.06
28	SCKNS2	.07	1.10	1.10	.53	.44	1.09	1.13	.66	.24	1.10	1.13	.57
29	SCKNS6	.22	.94	.93	1.26	.20	1.03	1.04	.89	.21	.98	.98	1.08
30	CCKES5	−.51	1.00	1.02	.97	−.62	1.08	1.04	.82	−.56	1.03	1.03	.89
31	CCKPO6	.26	.98	.97	1.10	.53	1.15	1.19	.52	.39	1.06	1.08	.77
32	KCSRPD4	.33	.98	.96	1.10	−.22	.96	.95	1.15	.07	.95	.94	1.21
34	CCKPO5	.44	1.01	1.01	.96	.91	1.10	1.15	.74	.66	1.07	1.09	.80
35	KCTLG4	1.46	1.02	1.01	.98	1.64	.97	.84	1.07	1.55	1.00	.95	1.02
36	SCKISS2	−.15	.95	.94	1.30	−.25	.94	.88	1.25	−.20	.94	.91	1.29
37	SCKISS5	−.61	.97	1.01	1.11	−1.24	.93	.80	1.14	−.88	.94	.92	1.16
38	SCKNS7	.33	.97	.97	1.11	.39	.98	.94	1.10	.36	.98	.96	1.10

Note.—Infit/outfit reported as MNSQ; Item = item difficulty; Dis = discrimination index. The first three letters in the item code represent the subdomain of MKT (e.g., CCK = common content knowledge). The next two or three letters correspond to the subcategory of knowledge in that domain (e.g., RPD = relative problem difficulty). The numeral at the end of the item code simply provides a unique identifier for the item within its subdomain and subcategory. See the test blueprint in Table 3 for the full names of the subdomains and subcategories. Items 6, 21, and 33 from the original test form were not included in the final scale.

samples. Given the small deviations from the expected within the fit statistics, it is not surprising the Rasch model accounted for a large portion of variance within the measure overall (77.1%), and for Sample 1 (79.0%) and Sample 2 (74.9%).

Discrimination Index

In the process of fitting the sample data to the Rasch model, the beginning steps within the analysis assume that all items have the same difficulty (e.g., 1.00) and fit the underlying model. Final derivations from this initial expected difficulty as influenced by the pattern of individual responses provide an indication of fit to the Rasch model. Linacre (2006) suggests a range for interpretable discrimination between 0.5 and 2.0, with values greater than 1.0 indicating highly discriminant items and values less than 1.0 as less discriminant. In other words, high-value discrimination items are more likely to be answered by teachers high in MKT than by teachers low in MKT. Low-discrimination items indicate less distinction between high or low MKT.

Based on pooled data from the two subsamples, Table 5 displays the discrimination estimate for each item. The highest discrimination values were found for items within the SCK and CCK trajectories across both samples and ranged from 1.25 to 1.48. Beyond highly discriminating within each sample, these questions demonstrate moderately high correlations to the overall measure.

Low discrimination values are more problematic, as these questions fail to differentiate between test takers of different ability levels. Across both samples, all items but one (Item 18; discrimination estimate = 0.45) were above the lower threshold of 0.5. That one item, however, demonstrates acceptable fit and was subsequently retained for examination within future data collection and analysis.

Person and Item Reliability and Item Separation

Person-separation reliability measures the degree to which the scale differentiates persons on the items. It is calculated by Winsteps as the ratio between the true person variance to the observed person variance and ranges from 0 to 1. Values greater than .80 are generally considered to indicate adequate reliability. Overall, the person reliability estimate fell slightly below the .80 cut point (.75). The lowest level of person reliability was found for Sample 1 (.66). With a person reliability of .79, the person reliability for Sample 2 is slightly below the preferred cutoff of .80, indicating adequate person-separation reliability for this group. The analogous Cronbach's alpha across all items and samples was .75 (Sample 1, $\alpha = .66$; Sample 2, $\alpha = .79$).

The person-separation index estimates the spread of individuals across the measure items and is calculated as the adjusted standard deviation divided by the error standard deviation. Values above 2.0 are indicative of adequate spread (Bond & Fox, 2007; Linacre, 2005). Less separation of persons across items was also seen within Sample 1 (1.40). Overall (1.74) and for Sample 2 (1.93), values were closer but still lower than the 2.0 cutoff.

Item-separation reliability quantifies how well a sample of participants can separate the items on the measure (Wright & Stone, 1999). It is calculated by Winsteps by dividing true item variance by the observed item variance (Bond & Fox, 2007)

and also ranges from 0 to 1. Item separation was excellent across the combined group as well as within both Samples 1 and 2 (.97–.98).

Person-separation indices indicate how efficiently a set of items is able to capture levels of skill within a sample (Wright & Stone, 1999). Person separations were good and suggested between five and eight levels of skill within the measure. Tables 6 and 7 display the item- and person-separation statistics.

Discussion

The goal of the work we report in this article was to create an assessment instrument that could be used efficiently at a large scale to measure MKT specific to the early elementary level. Informed by our review of the CCSS-M and the content of the PD programs and by external expert reviews of our items and test blueprint, we are confident that K-TEEM generates an interpretable score that corresponds to knowledge for teaching early elementary-grades mathematics. We think K-TEEM may be relevant for teachers of early elementary mathematics in general but is especially relevant for use with teachers working in environments where the CCSS-M (or similar standards) is a key feature in the school accountability system. The current K-TEEM scale focuses on the topics of number, operations, and algebraic thinking; it does not currently measure knowledge of other topics, such as measurement or geometry.

Through the field test of K-TEEM, the test administration and scoring procedures were determined to be an acceptable level of burden for both the teachers and the test administrators. The person- and item-reliability estimates appear to meet basic standards for educational and psychological measurement, although the discrepancy in some of the statistics between the two samples warrants further investigation and refinement of the items or scale.

The Rasch model produces a single score that lends itself well to typical models commonly used to investigate the effects of teacher PD interventions, such as multilevel analysis of covariance. The Rasch-based score may also be used to investigate associations between teachers' MKT and student learning outcomes—a link that has proven to be elusive in extant, large-scale studies. The Rasch model accounted for approximately 75% of the variance in the underlying construct of MKT and had excellent to acceptable levels of infit and outfit across both samples.

The assumption of unidimensionality in the Rasch model enabled us to obscure the potential distinctions among facets of knowledge in the subdomains of MKT in favor of defining their relationship to an underlying, singular construct. Because of the interrelatedness of the various facets of knowledge we were attempting to mea-

Table 6. Person-Separation Statistics

Sample	Average Measure	Infit (*SE*)	Outfit (*SE*)	Adjusted *SD*	RMSE	Separation	Reliability
Overall	.00	1.00 (.14)	1.00 (.23)	.70	.40	1.74	.75
Sample 1	−.17	1.00 (.14)	1.00 (.20)	.55	.39	1.40	.66
Sample 2	.43	1.00 (.14)	1.01 (.34)	.79	.41	1.93	.79

Note.—Overall, $n = 405$; Sample 1, $n = 206$; Sample 2, $n = 199$.

Table 7. Item-Separation Statistics

Sample	Average Measure	Infit (*SE*)	Outfit (*SE*)	Adjusted *SD*	RMSE	Separation	Reliability
Overall	.00	1.00 (.06)	1.00 (.10)	.90	.11	7.87	.98
Sample 1	.00	1.00 (.05)	1.00 (.09)	.90	.16	5.67	.97
Sample 2	.00	1.00 (.08)	.99 (.13)	.95	.17	5.61	.97

Note.—Overall, $n = 405$; Sample 1, $n = 206$; Sample 2, $n = 199$.

sure, the Rasch model seemed an appropriate analytic method at this stage, and the unidimensionality assumption appears to withstand some gentle scrutiny. Future work should examine the strength of the assumption of unidimensionality and possible alternative structures within K-TEEM.

Many researchers in mathematics education have been reluctant to use rigorous evaluation designs (e.g., randomized controlled trials) to measure the effect of educational interventions. This reluctance may, in part, be due to the limited availability of assessment instruments designed for large-scale studies that can pass muster with researchers in mathematics education. Indeed, poor alignment between the substantive content measured by a measurement instrument and the focus of an educational program is a major threat to internal validity of the results of a study. By design, the content of K-TEEM aligns with some of the most widely acknowledged findings in the corpus of research in early mathematics. By measuring the substantive content considered to be important by scholars in mathematics education, perhaps K-TEEM can provide a critically important tool to support the transformation toward rigorous evaluation designs becoming more common in mathematics education.

Methodological Considerations in MKT Item and Scale Development

In the following sections, we discuss a few particularly important considerations that may provide further insight into the content and structure of K-TEEM. We faced these decisions in the development of K-TEEM, but we think these considerations are generally applicable in the development of any assessment instruments intending to measure teacher knowledge.

Use of context in MKT items. It is common to include scenarios involving teachers and students in assessment items designed to measure MKT. This can be one aspect of the items in an MKT test that make it different from other types of tests of mathematical knowledge. Three of the released items in the Appendix use this technique. The technique is used partly in an attempt to demonstrate to the test taker that these questions or problems are relevant to the work they do as teachers. Used cleverly, it may also improve the interpretability of the underlying trait or ability the test is trying to measure by measuring knowledge or ability in a way that is situated in the setting or scenario in which the person might use that particular knowledge.

Figure 1 contains the original version of released Item 4 in the Appendix. We originally thought the context involving the word problem might provide support for the test takers to think about different ways to solve the problem. Through the

Mrs. Jones presented the following problem to her first grade class with intent to discuss the ways children might use their knowledge of number facts to help them solve problems.

> Jonah had 6 cars and 8 trucks in his toy vehicle collection. If Jonah displays his cars and trucks on a special shelf, how many vehicles will there be on the shelf?

Describe as many different ways as you can think that a child might solve this problem correctly.

Figure 1. A "solve many ways" item involving an unnecessary scenario.

cognitive interviews, we found that the introduction to the problem and the associated word problem within it drew a disproportionate amount of the test taker's attention and increased the length of time required to complete the item. We removed most of the context of this particular item, and the resulting item is released Item 4 in the Appendix. For many items, the context was determined to be an integral part of the item and contributed to the central idea in the item. In those cases, the context was retained. When the context was not necessary to serve the purpose of a given item, the context was removed to minimize reading time and cognitive load.

Verifying specialized and pedagogical content knowledge. One challenge in measuring SCK and PCK using test items scored dichotomously (as either correct or incorrect) is in determining types of knowledge that can be scored as correct or incorrect objectively and definitively. To protect the integrity of the interpretation of the score on the test as being free from bias with respect to certain epistemologies or theories of instruction, the items associated with the SCK, KCS, and KCT subdomains were required to meet at least one of two criteria: correctness by definition or by substantial empirical evidence. Released Item 1 in the Appendix conforms to these criteria, whereas released Items 2 and 3 were ultimately judged to be nonconforming to these criteria (and were consequently removed from the scale).

An example of items scored for correctness by definition are those that ask teachers to observe a child solving a problem and to select the name of the strategy used by the student. The names of strategy types and their corresponding definitions are provided in research literature published in refereed journals and highly credible summaries of those works (e.g., Carpenter et al., 1999; Sarama & Clements, 2009). To address differences in vernacular, a write-in option was made available if the available multiple-choice options did not include a correct answer in the form the test taker expected to see it. The write-in responses were subsequently reviewed by an adjudication committee to determine which ones were equivalent to the predetermined correct answer. For instance, Carpenter et al. (1999) use the term *direct modeling* to mean something very similar to what Sarama and Clements (2009) might call *concrete modeling*. In common teacher vernacular, teachers might call this same idea *concrete representation*. The adjudication process is used to monitor scoring and determine whether these responses might be considered to be synonymous (although the latter two were not observed in the field-test data).

For other items, we set an additional necessary condition of having sufficient empirical evidence gathered through studies published in refereed sources to support the judgment of correctness. One such category of knowledge in the KCS domain is that of relative word problem difficulty. The items in the relative problem

difficulty category of knowledge in the KCS domain and the veracity of their answers is determined by a synthesis of results in published data gathered between 1970 and 2000 and a recent replication of those findings using data gathered in the United States from 2013 to 2016 (Schoen, Champagne, & Whitacre, 2015).

Constructed- and selected-response item types. We began the development process with the perspective that open-ended items were inherently superior to items presented in a multiple-choice format for probing some facets of MKT, such as teachers' ability to interpret student solutions. Through our experiences in the scoring and cognitive interview processes, our confidence in the ability of open-ended items to reliably measure some important aspects of knowledge and ability decreased. We use released Item 2 in the Appendix as an example.

With released Item 2, we wanted to find out whether teachers would identify that students might use a relational thinking approach (Carpenter, Levi, Franke, & Zerinque, 2004) to quickly determine the value of the unknown quantity in the equation $46 + 27 = \_\_\_ + 26$. Some teachers offered responses such as "He added 1 to 46." This response leaves a lot of ambiguity and uncertainty with respect to the question of whether the teacher has insight into relational thinking. The item was removed from the final K-TEEM scale out of concerns about having to choose between misinterpreting responses or trivializing the scoring procedures. Of course, the underlying problem might not be in the item type but on some other aspect, such as clarification of the construct the item was intended to measure or a limitation of the question itself to yield insight into this type of knowledge.

Methodologists in educational and psychological measurement argue that, under certain conditions, selected-response (i.e., multiple-choice) items are superior to constructed-response items in their ability to reliably measure knowledge and ability (Downing, 2006). For future versions of K-TEEM, we intend to convert some of the constructed-response items into multiple-choice items by borrowing frequently observed ideas in the teachers' written responses—both correct and incorrect—to serve as selected-response options. Done well, we think this approach can simultaneously improve the measurement qualities of the item and the efficiency of the scoring process. While we have become more confident in the potential value and usefulness of selected-response items, we think it is critically important to find out how the target population actually responds to the questions and to use the respondents' exact words in the response options to maximize item reliability and validity for use with the population of interest.

Next Steps

We think that we have taken important first steps in the development of a valid and reliable way to measure MKT at the early elementary level. We also expect that further development and investigation will provide important insight into both the underlying construct that we are trying to measure and the appraisal of the validity of use for its intended purpose. For instance, testing whether K-TEEM is sufficiently sensitive to detect group differences in MKT will be important for validation purposes.

Other natural next steps will involve multidimensional models (e.g., factor-analytic models, multidimensional models based on item response theory) and an analysis of differential item functioning (DIF). It will be important to perform a DIF

analysis to investigate whether items are biased with respect to specific PD programs or are invariant with respect to repeated measurement or individual characteristics of teachers. A DIF analysis and (possible) respecification of the test composition based on the results of the DIF analysis could yield insight into the discrepancies in item difficulty, model fit, and reliability observed between Samples 1 and 2. An investigation of multidimensionality may yield empirical insights into separation between theorized subdomains within the MKT construct. Both procedures will require larger samples of teachers to provide sufficient statistical power.

Surprisingly few large-scale, empirical findings support the claim of a correlation between teacher knowledge and student learning in mathematics or in other subject areas. Given the overwhelming agreement among scholars with the claim that the association should exist, further investigation of these associations is very important. We have student achievement data for half of the teachers in the current sample, and another particularly important next step will involve investigation of whether teachers' scores on the K-TEEM scale can predict student achievement or learning gains.

Conclusions

Developing an assessment instrument to meet standards in educational and psychological measurement is a humbling experience that requires tremendous attention, effort, and expertise. Ultimately, our goal will be to offer K-TEEM to the research and evaluation community for use. While it is not yet perfect, we think K-TEEM may fill an important gap in the set of tools available to researchers and evaluators for the purpose of investigating the associations among teacher MKT, PD, student learning, and other factors of interest.

Appendix

Released Items from the Questionnaire Development Process

Released Item 1 (Multiple Choice)

Sequence the three problems that follow from least difficult to most difficult for most first graders at the beginning of the year to solve correctly.

Note: You may assume that the students can have the problems read aloud as many times as needed and that they have the option to use paper and pencil or manipulatives.

Problem A	Problem B	Problem C
The candy bowl has 5 peppermint candies and 14 butterscotch candies. How many more butterscotch candies are there than peppermint candies?	There were some candies in the bowl. Anna came and put 9 new candies in the bowl. Now the bowl has 14 candies. How many candies were in the bowl before Anna came?	The candy bowl was filled to the top with 14 candies. Anna grabbed 5 candies out of the bowl to share with her friends. How many candies are in the bowl now?

a. A, C, B
b. B, C, A
c. C, A, B
d. C, B, A

Correct answer: *c*

Released Item 2 (Open-Ended Response)

$$46 + 27 = \square + 26$$

Mr. Johnson presented this equation to his first-grade class. Without writing anything, a student quickly called out that the missing number is 47.

What is the most likely explanation for how the student generated the correct answer so quickly?

Scoring: Credit given if explanation describes use of relational thinking.

Released Item 3 (Multiple Choice)

Ms. Reynolds believes several of her first-grade students are ready to progress from using cubes or pictures to represent all of the quantities in Join (or Add To) Result Unknown problems to using a counting on strategy. Which of the problems below has numbers that are most likely to nudge these students to use counting on instead of modeling all quantities with objects or pictures?

Choose the one best answer:

a. Jon had 12 stickers in his collection. His grandma gave him 9 more stickers. How many stickers does Jon have now?
b. Jon had 6 stickers in his collection. His grandma gave him 7 more stickers. How many stickers does Jon have now?
c. Jon had 5 stickers in his collection. His grandma gave him 8 more stickers. How many stickers does Jon have now?
d. Jon had 23 stickers in his collection. His grandma gave him 2 more stickers. How many stickers does Jon have now?

Correct answer: *d*

Released Item 4 (Open-Ended Response)

Describe as many ways as you can think of that a child might use number-fact knowledge to correctly find the sum of 6 + 8.

Please provide a detailed description or notation of the steps in each strategy, using the specific numbers from the problem and making clear how the answer is determined.

Strategy 1:

Strategy 2:

Strategy 3:
Strategy 4:
Strategy 5:
Strategy 6:

Scoring: To receive credit for this item, description or notation of four valid and distinct strategies for using number fact knowledge to solve 6 + 8. Specific numbers in the problem must be included.

Notes

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305A120781 to Florida State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Robert C. Schoen is the associate director of the Florida Center for Research in Science, Technology, Engineering, and Mathematics (FCR-STEM) in the Learning Systems Institute at Florida State University. Wendy Bray is research faculty at the FCR-STEM in the Learning Systems Institute at Florida State University. Christopher Wolfe is an assistant professor of psychology in the Social Sciences Department at Saint Leo University. Amanda Tazaz is an associate in research at the Learning Systems Institute at Florida State University. Lynne Nielsen is an assistant professor of mathematics education in the College of Education, Curriculum, Instruction, and Leadership at Louisiana Tech University. Correspondence concerning this article should be addressed to Robert C. Schoen, Learning Systems Institute, Florida State University, 4600 C University Center, Tallahassee, FL 32312. E-mail: rschoen@lsi.fsu.edu.

1. The lead author of this article completed the training and orientation required to use the LMT items and scales.

2. This list of activities and learning goals should not be interpreted as comprehensive. Rather, it is intended merely to describe some of the central activities and learning goals in the program.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D., & Marais, I. (2014). Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement*, **38**(6), 432–499.

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.

Bodenhausen, J., Denhart, N., Gill, A., Kaduce, M., & Miller, M. (2014). *Thinking mathematics K–2 Common Core edition.* Washington, DC: American Federation of Teachers.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Borko, H., & Putnam, R. T. (1996). Learning to teach. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 673–708). New York: Macmillan Library Reference.

Bush, W. S., Ronau, R., Brown, T. E., & Myers, M. H. (2006, April). *Reliability and validity of diagnostic mathematics assessments for middle school teachers.* Paper presented at the annual meeting of the American Educational Association, San Francisco.

Campbell, P. F., Rust, A. H., Nishio, M., DePiper, J. N., Smith, T. M., Frank, T. J., . . . Choi, Y. (2014). The relationship between teachers' mathematical content and pedagogical knowledge,

teachers' perceptions, and student achievement. *Journal for Research in Mathematics Education*, **45**(4), 419–459.

Carlisle, J. F., Kelcey, B., Rowan, B., & Phelps, G. (2011). Teachers' knowledge about early reading: Effects on students' gains in reading achievement. *Journal of Research on Educational Effectiveness*, **4**(4), 289–321.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction.* Portsmouth, NH: Heinemann.

Carpenter, T. P., Levi, L., Franke, M. L., & Zerinque, J. K. (2004). Algebra in elementary school: Developing relational thinking. *ZDM*, **37**(1), 53–59.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, **26**(1), 1–22.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Mahwah, NJ: Erlbaum.

Fennema, E., Carpenter, T. P., Levi, L., Franke, M. L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction—A guide for workshop leaders*. Portsmouth, NH: Heinemann.

Fennema, E., & Franke, M. L. (1992). Teachers' knowledge and its impact. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 147–164). New York: Macmillan.

Fuson, K. (1992). Research on whole number addition and subtraction. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning.* Reston, VA: National Council of Teachers of Mathematics.

Garet, M. S., Heppen, J., Walters, K., Smith, T., & Yang, R. (2016). *Does content-focused teacher professional development work? Findings from three Institute of Education Sciences studies* (NCEE Evaluation Brief 2017-4010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Goldsmith, L. T., Doerr, H. M., & Lewis, C. C. (2014). Mathematics teachers' learning: A conceptual framework and synthesis of research. *Journal of Mathematics Teacher Education*, **17**(1), 5–36.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, **15**(3), 309–334.

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institutes. *Journal for Research in Mathematics Education*, **35**(5), 330–351.

Hill, H. C., Ball, D. L., Blunk, M., Goffney, I. M., & Rowan, B. (2007). Validating the ecological assumption: The relationship of measure scores to classroom teaching and student learning. *Measurement*, **5**(2/3), 107–118.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, **42**(2), 371–406.

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

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts? In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 111–155). Charlotte, NC: Information Age.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Linacre, J. M. (2005). *Winsteps Rasch measurement* (Version 3.58.1) [Computer program]. Chicago: Winsteps.

Linacre, J. M. (2006). Data variance explained by Rasch measures. *Rasch Measurement Transactions*, **20**(1), 1045–1054.

Linacre, J. M. (2016). Winsteps (Version 3.7) [Rasch measurement computer program]. Beaverton, OR: Winsteps.

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

Ma, L. (1999). *Knowing and teaching elementary mathematics*. Mahwah, NJ: Erlbaum.

Moats, L. (2009). Still wanted: Teachers with knowledge of language. *Journal of Learning Disabilities*, **42**(5), 387–391.

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

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington DC: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

Nesher, P., Greeno, J. G., & Riley, M. S. (1982). The development of semantic categories for addition and subtraction. *Educational Studies in Mathematics*, **13**(4), 373–394.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded edition, with a foreword and afterword by Benjamin D. Wright). Chicago: University of Chicago Press. (Original work published 1960)

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct modeling approach. *Journal of Educational Psychology*, **103**(1), 85–104.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, **6**(1), 43–74.

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**(1), 180–192.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children.* New York: Routledge.

Schoen, R. C., Champagne, Z. M., & Whitacre, I. (2015, November). *Re-examining the validity of word problem taxonomies in the Common Core era.* Paper presented at the annual conference of the North American Chapter of the International Group for the Psychology of Mathematics Education, East Lansing, MI.

Schoen, R. C., LaVenia, M., Champagne, Z., & Farina, K. (2016). *Mathematics Performance and Cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2014* (Report No. 2016-01). Tallahassee: Florida State University, Learning Systems Institute. doi:10.1725/fsu.1493238156

Schoen, R. C., LaVenia, M., Champagne, Z. M., Farina, K., & Tazaz, A. (2016). *Mathematics Performance and Cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2015* (Report No. 2016-02). Tallahassee: Florida State University, Learning Systems Institute. doi:10.17125/fsu.1493238666

Seidel, H., & Hill, H. C. (2003). *Content validity: Mapping SII/LMT mathematics items onto NCTM and California standards.* Ann Arbor: University of Michigan, School of Education.

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

Slavin, R., & Madden, N. A. (2011). Measurement inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, **4**(4), 370–380.

Smith, P. S., & Banilower, E. R. (2006, April). *Measuring teachers' knowledge for teaching force and motion concepts.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco.

Sowder, J. T. (2007). The mathematical education and development of teachers. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 157–223). Charlotte, NC: Information Age.

Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole numbers concepts and operations. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning.* Charlotte, NC: Information Age.

Wheeler, G. D. (2010). *Assessment of college students' understanding of the equals relation: Development and validation of an instrument* (Unpublished doctoral dissertation). Utah State University, Logan.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues and Answers Report, REL 2007-033). Washington, DC: U.S. Department of Education, Institute of Education Sciences.