2016 NCTM RESEARCH CONFERENCE
April 11-13 • San Francisco

Paper Title: Measuring Primary Grades Teachers' Mathematical Knowledge for Teaching
Author(s): Wendy S. Bray and Robert C. Schoen
Session Title: Measuring Primary Grades Teachers' Mathematical Knowledge for Teaching
Session Type: Brief Research Report
Presentation Date: April 13, 2016
Presentation Location: San Francisco, California

Measuring Primary Grades Teachers' Mathematical Knowledge for Teaching

Wendy S. Bray
Robert C. Schoen
Florida Center for Research in Science, Technology, Engineering, and Mathematics (FCR-STEM) at Florida State University

Deborah Ball and her colleagues have coined the term *Mathematical Knowledge for Teaching* (MKT) to describe facets of knowledge used in the practice of teaching mathematics (Hill, Rowan, & Ball, 2005). While many professional development initiatives are designed with intention to increase various facets of teachers' MKT (Sowder, 2007), few have undertaken rigorous evaluation to empirically measure change in teachers' MKT as a result of specific professional development initiatives (Gersten, Taylor, Keys, Rolfhus, & Newman-Gonchar, 2014; National Mathematics Advisory Panel, 2008). One reason for this is that research focused on development of valid and reliable measures of MKT is a fairly new direction in the field of mathematics education (Hill, Sleep, Lewis, & Ball, 2007), so there are a limited number of instruments available. In particular, there is a critical need for additional development work on measures that focus on the MKT involved in teaching primary grades mathematics (i.e., K–2) (Hill & Ball, 2004; Seidel & Hill, 2003).

To this end, we have developed a measure of MKT that comprises subject matter closely aligned to the work of teaching number, operations, and equality in the primary grades. This paper will share our development process, including how we have conceptualized and specified MKT for scale development. We will also report results from field tests involving two iterations of the measure and take a close look at how our development process was used to strengthen particular items. Finally, we will discuss what we have learned from our data and the process of scale development.

Conceptualization of MKT for Scale Development

In the mid-1980s, Lee Shulman and colleagues theorized subject matter knowledge to be comprised of both *content knowledge* and *pedagogical content knowledge* (Shulman 1986, etc.). Building on Shulman's work, Ball and colleagues further elaborated the constructs of content knowledge and pedagogical content knowledge for the subject of mathematics (Ball, Thames, and Phelp, 2008). In their framework of Mathematical Knowledge for Teaching (MKT), subdomains that comprise the larger domain of MKT are identified. Sub-domains elaborating facets of content knowledge include *common content knowledge* (CCK), *specialized content knowledge* (SCK), and *horizon content knowledge* (HCK). Sub-domains elaborating facets of pedagogical content knowledge include *knowledge of content and students* (KCS), *knowledge of content and teaching* (KCT), and *knowledge of content and curriculum* (KCC).

In our work developing a scale to measure primary grade teachers' MKT with focus on K-2 mathematics content, we developed items with intention to reflect the sub-domains of CCK, SCK, KCS, and KCT. Each of these four subdomains will now be described with particular attention to the specification of types of items included on our scale to measure relevant knowledge related to the work of teaching number, operations, and equality concepts of focus in the primary grades:

Common Content Knowledge (CCK) is mathematical knowledge and skill that is useful in multiple settings and therefore is not unique to the work of teaching (Ball, Thames, & Phelps, 2008). For example, persons with strong CCK can approach mental computation tasks in multiple ways, employing flexible strategies depending on the quantities involved in specific calculations. Our working definition of CCK also includes knowledge of accepted meaning of mathematical vocabulary and conventions of notation. The items developed for our scale focus on three types of CCK that we view as particularly relevant to teaching primary grades mathematics: 1) knowledge of the mathematical meaning of the equal sign and related conventions of notation, 2) ability to identify properties of operations by name, in concept, and in context of a solution to a specific mathematics problem, and 3) knowledge of how to solve arithmetic problems in different ways.

Specialized Content Knowledge (SCK) is knowledge of mathematics that is uniquely useful in teaching (Ball, Thames, & Phelps, 2008). Teachers use SCK to make sense of the mathematical details of the various correct and flawed approaches students take as they solve problems. We also include in the realm of SCK knowledge of professional vernacular related to mathematics teaching and learning that is found in primary grades mathematics standards documents. The items included on our scale focus on three types of SCK that we view as particularly relevant to teaching mathematics in the primary grades: 1) ability to analyze student solutions to arithmetic problems and identify errors or make sense of correct solutions that may be mathematically valid yet use non-standard methods or notation, 2) knowledge of names of word problem types associated with different semantic structures in accordance with the Common Core State Standards for Mathematics or the typologies presented in Cognitively Guided Instruction, 3) ability to identify acknowledged mathematical strategy names (i.e., such as *compensation* or *standard algorithm*) when presented with student work or video of a student solving a problem.

Knowledge of Content and Students (KCS) is “knowledge that combines knowing about students and knowing about mathematics” (Ball, Thames, & Phelps, 2008, p. 401). A teacher with strong KCS holds knowledge of how students tend to think about and approach particular mathematical concepts and types of problems. Drawing on the corpus of research focused on how primary grades children think about mathematical ideas (for example, Carpenter, Hiebert, & Moser, 1981; Carpenter & Moser, 1984; Carpenter, Moser, & Romberg, 1982; Hiebert, 1982; Turner & Celedon-Pattichis, 2010), our scale includes items designed to tap into two types of KCS knowledge: 1) knowledge of the relative difficulty of word problems with different semantic structures, and 2) ability to predict how young children are most likely to approach word problems with different semantic structures.

Knowledge of Content and Teaching (KCT) is knowledge of how particular teaching strategies and instructional sequences can be used in service of helping learners deepen understanding of specific mathematical ideas (Ball, Thames, and Phelps, 2008). One important aspect of KCT is being able to design or choose mathematical tasks that match particular instructional goals. For example, primary grades teachers with strong KCT are able to pose a ‘just right’ word problem to provoke some children to use *counting on from larger* instead of *counting on from first* when solving an addition problem. Our scale includes items designed to probe the facet of KCT used to select word problems in the service of specified instructional goals that are important to primary grades mathematics.

Instrument Development

Our immediate motivation for developing this instrument was a desire to measure the MKT of primary grades teachers involved in two different professional development programs, both focused on deepening primary grades teachers' understanding of mathematics and children's mathematical thinking related to number, operations, and equality. Therefore, we began our instrument development process by analyzing the particular aspects of MKT that the professional development programs set out to increase and its overlap with student learning goals defined by the Common Core State Standards Mathematics (CCSSI; National Governors Association Center for Best Practices & Council of State School Officers, 2010). Using these parameters, an item blueprint was developed (see Table 1) to operationalize the identified sub-domains of MKT and guide drafting of items.

Table 1
Initial Item Blueprint

Sub-domain of MKT	Categories of Items
Common Content Knowledge	• Meaning of the Equal Sign and Related Notation • Properties of Operations • Solve Problems in Many Ways
Specialized Content Knowledge	• Evaluating the Validity or Generalizability of Student Strategies[a] • Naming Student Strategies • Naming Word Problem Types • Writing Word Problems[a]
Knowledge of Content and Students	• Predicting Students Strategies • Relative Problem Difficulty • Matching Strategies and Problems[a]
Knowledge of Content and Teaching	• Selecting Word Problems in Service of Specific Instructional Goals

Note.
[a]These categories were reconceptualized or dropped later in the development process.

After the initial drafting of items, we vetted and refined the item bank by seeking feedback from experts and conducting cognitive interviews with teachers. First, we sought feedback on draft items from experienced classroom teachers, teacher educators, and other experts in mathematics and mathematics education. In particular, we asked these experts to identify what they thought each item was measuring and comment on the clarity and validity of items. This initial round of feedback led to some revision and elimination of items.

Second, we conducted two rounds cognitive interviews with elementary teachers in which teachers completed draft items one-by-one and verbalized their thought processes as they interacted with and reflected on each item (Desimone & LeFloch, 2004). Five teachers participated in round one of cognitive interviews, and six teachers participated in round two. All cognitive interviews were audio-recorded to facilitate review and subsequent analysis. During cognitive interviews, interviewers (who were all members of the instrument development team) asked probing questions to gain insight into teachers' interpretation of items, what teachers considered as they determined their answers, and teachers' confidence in their answers. Interviewers also made observations about the amount of time and cognitive effort required for each item as well as the respondent's affective reaction to each item. After each round of cognitive interviews, careful analysis and comparison across interviews led to items being revised or eliminated, and a few new items were created.

This process resulted in a scale of 40 items consisting of 30 multiple-choice items, 3 fill-in-the-blank items, and 7 constructed-response items. These items were then transferred into an on-line platform and piloted in Spring 2014 in the context of a large-scale field test. Field test participants were primary grades teachers (K-2) and support personnel involved in unrelated randomized controlled trials of two different programs of mathematics professional development (n = 413): 1) Cognitively Guided Instruction (CGI), developed and taught by Teacher Development Group, or 2) Thinking Mathematics, developed and taught by the American Federation of Teachers. Following the field test, criteria were used to dichotomously score fill-in-the-blank and constructed-response items, and statistics were generated using Rasch models for the full scale and individual items (Rasch, 1980). Five items were dropped during the data analytic process due to poor item statistics or item infit or outfit statistics generated during the analysis of the 2014 field test data, yielding a set of 35 items in the final 2014 scale. Following the 2014 field test, items were further refined, and a revised scale of 37 items was piloted in Spring 2015. We refer to the assessment as Knowledge for Teaching Early Elementary Mathematics (K-TEEM). To allow for versioning, we reference the scale used in the 2014 field test as the 2014 K-TEEM and the scale used in the 2015 field test as the 2015 K-TEEM.

The 2015 field test sample was comprised of elementary teachers and support personnel (*n* = 636) participating in three different randomized controlled trials of teacher professional development. Similar to the 2014 field test, all of these teachers had applied to participate in a one- or two-year professional development program in mathematics and were working as a certified teacher in one of 35 different school districts in a single state. About half of these teachers had also been in the treatment or the control groups for the two previously mentioned studies and had completed the 2014 version of the test one year prior. An additional 271 teachers completed the assessment as part of their pretest prior to the first year of starting the CGI program. These additional 271 teachers had not completed any of the CGI or Thinking Mathematics professional development programs. Thus, approximately 200 of the teachers who completed the test in 2015 had completed either one or two years of the CGI or Thinking Math professional development programs. Approximately 400 of the 636 teachers had not yet participated in either of these programs. Following the 2015 field test, statistics were again generated using Rasch models for the full scale and individual items for an analytical sample of 36 items.

Table 2 provides the final item blueprint, summarizing the categories of items represented in the scale and the final number of items in each category after data analysis based on the Spring 2014 and Spring 2015 field tests.

Table 2

Blueprint of Items by MKT Sub-domain and Sub-category included in Analyses

Sub-category of Items by Sub-domain of MKT	Number of Items 2014	Number of Items 2015
Common Content Knowledge		
Meaning of the Equal Sign and Related Notation	5	5
Properties of Operations	4	5
Solve Problems in Many Ways	2	2
Specialized Content Knowledge		
Interpreting Student Strategies	4	4
Naming Student Strategies	4	5
Naming Word Problem Types	5	4
Knowledge of Content and Students		
Predicting Student Strategies	3	3
Relative Problem Difficulty	4	4
Knowledge of Content and Teaching		
Selecting Word Problems in Service of Specific Instructional Goals	4	4
Total Items:	35	36

Findings

Our findings will now be presented in two parts. First, we will present results on the scales used in field tests conducted in 2014 and 2015. Then, we will trace the development of 3 items to offer insight into how our item development process influenced the evolution of items.

Field Test Results

Rasch model output data were used to determine the extent to which items had good discrimination estimates and fit the model. The assumption of unidimensionality, which is essential for analyses grounded in item response theory (IRT), was reasonably well supported by the model data. In both years, the Rasch model accounted for more than 75% of the total variance in the sample.

Model fit. Items with poor infit or outfit statistics were removed prior to the final models in both years. Items were considered misfit if the MNSQ estimates were either less than 0.6 or greater than 1.4. Low values of MNSQ may be indicative of redundancy with other items, while high values may indicate items out of sync with other items in the measure (Linacre, 2005). As a result of the infit/outfit analysis, four items were removed in 2014, and one item was removed in 2015. As a result, the analytic sample was comprised of a total of 35 items for the 2014 field test and 36 items for the 2015 field test.

Reliability. Based on the sample of 413 teachers in 2014, the item-separation reliability was .75, and the person-separation reliability was .98. Based on the sample of 636 teachers in 2015, the item-separation reliability was .73, and the person-separation reliability was .99.

Concurrent validity. In effort to establish concurrent validity with another measure of teacher MKT, 66 teachers who took the 2014 K-TEEM in March 2014 also completed an additional assessment in June 2014. The additional assessment consisted of a 23-item scale created from the LMT item bank. Each of the 23 items addressed topics in the domain of number, operations, and equality. All 23 items involved only whole numbers (i.e., no fractions or decimal fractions) and involved specific numbers (i.e., did not use letters as unknown or generalized variables).

The LMT measure was also submitted to the Rasch model for comparison on the underlying unidimensional construct of teacher mathematics knowledge. The 23-item scale constructed from items drawn from the LMT item bank was not an excellent fit to the Rasch model. Person reliability for the 23-item scale was estimated at .64, and item reliability was estimated at .90. Because the SII/LMT items are considered valid, we retained all items for comparison to our scale. The overall correlations between the two measures was statistically significant at a threshold of $p<.05$ with a correlation coefficient of $r^2=.563$. This suggests a moderate level of overlap but that there may be differences in the type of knowledge that the two instruments are measuring.

Item Development Trace for Select Items

To illustrate how elements of our development process influenced individual items, we will now present an in-depth account of the evolution of three items. These items were selected, because they are items that have undergone significant revision and illuminate what we consider to be important lessons learned through the development and field testing process. While most items on our scale were revised through the development process, we want to be clear that not every item on our scale has been impacted to as great an extent as the ones we will share now.

The case of CCK.SMW.5. Item CCK.SMW.5 is one of a set of items designed to tap respondents' knowledge of how an arithmetic problem can be *solved in many ways* (SMW). This particular item is focused on probing knowledge of different ways a person can use related facts to solve a basic subtraction fact. Cognitive interview data led to significant revision of the content and structure of this item. Field test results led to additional revision of the item and further considerations related to measuring this construct. Figure 1 presents the item as it was originally drafted before cognitive interviews and as it appeared on the scale used in the 2014 field test.

Cognitive interviews on the original draft of CCK.SMW.5 revealed multiple findings that led to further revision of the item. First, multiple respondents appeared to get bogged down in reading and re-reading the SMW item directions and word problem presented in the original

2item. As respondents attended to those details, some glossed over the intended directive of the prompt – to identify different ways *number facts* could be used to solve 17 − 9 = ___. Instead, these respondents demonstrated how students might use strategies such as pictures or counting on, which was out of sync with the directive in the item. Given that our intent was to measure respondents' knowledge of how to solve arithmetic problems in many ways using known number facts and relationships, the effort spent making sense of the word problem and task verbiage was deemed an undesirable use of respondents' energy and attention. Additionally, we really wanted this item to hone in on whether the respondent could generate many different ways to use related facts to solve the problem, because we hypothesized that people who can think of many different ways to solve the problem have higher levels of MKT. A decision was made to streamline the item such that the directive to respondents was much shorter, and the contextualized problem was removed and replaced with a computation task. We also replaced the term *number facts* with the term *basic facts*, because some cognitive interview respondents identified the later term as more familiar to them.

Figure 1. CCK.SMW.5 Before and After Vetting Process

Original Draft Item	Revised Item on 2014 Field Test
Mrs. Fox selected the following problem with intent to have her second grade students discuss how knowledge of number facts might be used to solve the problem. Christina's pencil box was filled with 9 sharpened pencils and some unsharpened pencils. If there were 17 pencils in the pencil box, how many of the pencils were unsharpened? Describe as many ways as you can think of that a child might use **number fact knowledge** to solve the problem correctly.	Describe a variety of different ways a student could use **basic facts** to correctly solve 17 − 8 = ____. Please be specific in your descriptions. Strategy 1: Strategy 2: Strategy 3: Strategy 4: Strategy 5: Strategy 6:

Additionally, cognitive interviews suggested that the original item's reference to the setting of a "second grade class" led some respondents to situate the problem in their own interpretations of what is or should be taught in second grade. For instance, one respondent verbally commented that one could solve 17 − 9 by thinking 17 – 10 + 1 but that she wouldn't expect a second grader to do that. Rather, she asserted that second graders are usually taught to count up from 9 to 17 or to think, "9 plus what equals 17?" Consequently, the respondent did not include the derived fact strategy involving 17 − 10 in her written response to the item, even though this strategy was clearly part of her personal knowledge base. Since the intended focus of this item was on measuring content knowledge rather than knowledge of students or pedagogy, the item was revised to remove the reference to a particular grade level and focus entirely on the mathematics. In general, cognitive interviews have taught us to become much more sensitive to the effect the invocation of a particular grade level has on respondents' thinking when they consider their responses to items.

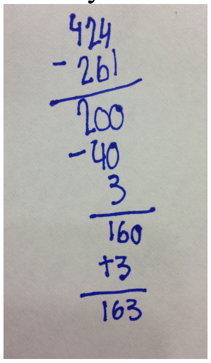
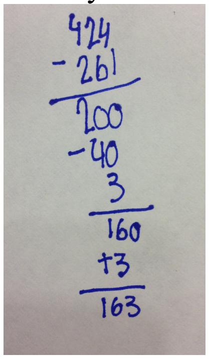
Another finding revealed through cognitive interviews of this SMW item was that, even though the item stated, "Describe as many ways as you can think of…," most respondents stopped working on the item after listing 3 ways. However, when prompted in the interview setting, some respondents were able to generate additional answers. In efforts to communicate our desire for respondents to generate as many ways as they could muster, the format of the on-line presentation of the item was revised from a single large textbox to a numbered list of 6 open field textboxes.

Following the 2014 field test, responses to CCK.SMW.5 were hand scored by at least two or three individual members of the development team. Discrepancies among scorers were resolved by the full team of five people. Responses received credit for the item if they provided 4 or more specific descriptions of distinct and valid strategies for using basic facts to correctly solve $17 – 8 = x$. Rasch model data suggested that while this item had acceptable fit to the scale as a whole, the item had one of the lowest percent correct statistics on the measure. During the hand scoring process, it was observed that many respondents were not receiving credit for the item because their responses did not provide enough evidence that the respondent understood the strategy of focus. For instance, many respondents listed names for strategies such as "Make Ten Strategy." Our criteria for scoring did not view this naming of strategies to be strong enough evidence that the respondent held knowledge of how the strategy would work for the specific numbers in the problem $17 – 8 = x$. But, of course, we expect that some of these respondents did hold knowledge to provide specific strategy details. In efforts to address this validity concern, we made additional revisions to the item ahead of the 2015 field test. We revised, "Please be specific in your descriptions" to, "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."

While responses scored as part of the 2015 field test suggested that our revised instructions did make more clear what we were looking for, we have become concerned about a different issue. Many respondents did not receive credit for a correct answer on this item, because they listed three or fewer strategies in their response (and the criteria to receive credit was to list at least four distinct and valid strategies for using basic facts to correctly solve $17 – 8 = x$). We wonder whether the item may require more time and effort than some teachers are willing to give to a single item. Particularly since the respondents in our field tests completed the survey on their own time through the on-line survey platform, we think it likely that some respondents hold knowledge of how to solve $17 – 8 = x$ in more ways than were evidenced by their item responses. As a result of this conjecture, we are have recently been working on developing an alternative item format for measuring knowledge of how to solve arithmetic problems in many ways that utilizes a multiple choice format.

The case of SCK.ISS.1. Item SCK.ISS.1 is one of a set of items aiming to measure the extent to which respondents can *interpret student strategies* (ISS) for solving arithmetic problems, with focus on strategies that are non-standard or flawed. Of focus in this particular item is interpreting a non-standard (but quite common) strategy for multi-digit subtraction in which a student has utilized negative numbers or reserved quantities to subtract later instead of employing standard re-grouping procedures (see item presented in Figure 2).

Figure 2. SCK.ISS.1 as Presented on Field Tests in 2014 and 2015

2014 Field Test Item	2015 Field Test Item
Melody, a second grade student, solved the following problem: Melody has 424 Legos. It takes 261 Legos to build a fort. How many Legos will she have left after building the fort? Melody solved the problem like this:  As you can see, Melody wrote –40 between the 200 and the 3 in the middle of her written work. Why do you think she wrote –40 there?	Melody, a second grade student, solved the following problem: Melody has 424 Legos. It takes 261 Legos to build a fort. How many Legos will she have left after building the fort? Melody solved the problem like this:  As you can see, Melody wrote –40 between the 200 and the 3 in the middle of her written work. Select the most likely explanation for why she wrote the –40 there. a. She made a mistake in her subtracting in the tens column. She incorrectly subtracted 60 – 20 to get 40 instead of regrouping. b. She rounded to numbers she was comfortable working with. She knew 60+40=100, so she found ways to work with that known fact. c. She subtracted 200, but temporarily ignored 24 of the 424. She then subtracted 40 of the 61 remaining in 261. This meant she then had 21 to subtract from 24 which gave her the 3. d. She wrote –40, because she knew the difference between 200 and 160 was forty. e. She subtracted 20 from the 20 in the tens column and wrote -40 as a reminder that she still needed to subtract 40.

Feedback from expert reviewers and data from cognitive interviews prior to the 2014 field test suggested that this constructed-response item would be difficult for respondents but yielded only minor revisions. In cognitive interviews, teachers consistently appeared to know what the item was asking; but the strategy of focus was novel to some, and sometimes teachers struggled to successfully interpret how the strategy worked.

Following the 2014 field test, responses were scored dichotomously using established criteria. In order to be considered correct, a response needed to offer a viable explanation for how the −40 makes sense in the context of Melody's solution. Specifically, respondents needed to either acknowledge the "negative" 40 or the need to notate that, after subtracting 20 in the tens column, 40 still needed to be subtracted. As previewed by the cognitive interviews, this item yielded a low percentage of correct responses. The item also had poor infit and outfit, so a decision was made to drop it from the final 2014 K-TEEM scale.

However, through the hand scoring process, it was noted several common responses that were repeatedly observed among the incorrect responses. We decided to select some of these incorrect responses and retool the item utilizing a multiple-choice format (see 2015 version in Figure 2). We conjectured that the multiple choice version of the item would reduce the difficulty and possibly help the item have better fit with the rest of the scale. The retooling of this item to multiple-choice would also carry the benefit of decreasing the number of items that required hand scoring, and it might make the item less time consuming for respondents.

We used the multiple-choice version of the item in the 2015 field test. The Rasch model once again indicated that the estimate for the item difficulty was very high, and the estimate for discrimination was far below the preferred .50 minimum. As a result, the item was dropped again from the final scale.

While this item still needs some work, we think shifting to a multiple-choice format progressed the item in a positive direction. In the next iteration of this item, we have reduced the number of choices to four such that the two distractor choices most frequently selected by respondents in 2015 were retained and the correct answer was modified slightly to be less wordy. We also eliminated two responses that were not frequently selected and added one new distractor choice that was also inspired by constructed responses from the 2014 field test.

The case of KCS.PS.2. Item KCS.PS.2 is one of a set of items designed to measure respondents' ability to *predict student strategies* (PS) given some information about the students. Multiple items in this category are grounded in the finding from research on children's mathematical thinking that, for some word problem types, young learners tend to devise strategies that closely match the semantic structure of a word problem (Carpenter, Fennema, Franke, Levi, & Empson, 2015; Carpenter & Moser, 1982; Hiebert, 1982; Nesher, Greeno, & Riley, 1982). KCS.PS.2 was intended to probe respondents' sensitivity to the fit between semantic structure of a story problem, the position of the unknown variable, and student strategy for a word problem involving a separating action with the result unknown. Expert review and cognitive interview data led to important revisions to the content of this item, and subsequent field test results prompted a rethinking of the set of 'predicting strategies' items. Figure 3 presents the item as it was originally drafted and as it appeared on the scale used in the 2014 field test.

Figure 3. KCS.PS.2 Before and After the Vetting Process

Original Draft Item	Revised Item on 2014 Field Test
This problem was posed to a class of first and second grade students: Kevin and his mom ordered a pizza with 8 slices. They ate 5 of the slices. How many slices of pizza did not get eaten? Of the choices below, which is a **counting strategy** that children who generally follow the structure of a story problem are **most likely** to use? Choose the one best answer: a. A student put up 8 fingers. Then she counted, “1, 2, 3, 4, 5” as she put down one finger at a time for each count. She counted the 3 fingers that were still up to get the answer. b. A student touched her head and said, “5.” Then she counted, “6, 7, 8,” extending one finger for each count. She then looked at her 3 extended fingers and said, “The answer is 3.” c. A student counted out a set of 8 cubes. Then he removed 5 cubes from the set, counting by ones, “1, 2, 3, 4, 5.” He counted the remaining cubes to find his answer, 3. d. A student said “8, 7, 6, 5, 4,” extending one finger for each count. Then he said, “There are 3 slices left.” e. A student wrote on her paper 8 – 5 = 3. She said, “The answer is 3.”	This problem was posed to a class of first and second grade students who usually use strategies that follow the structure of a story problem: Marco and his mom ordered a pizza with 12 slices. They ate 3 of the slices at the pizzeria and took the rest home. How many slices of pizza did they take home? Of the choices below, which one describes the strategy that these students are most likely to use? a. A student said, “12, 11, 10,” extending one finger for each count, and paused. She then looked at her fingers and said, “Nine slices are left.” b. A student said, “3” and paused. She then counted, “4, 5, 6, 7, 8, 9, 10, 11, 12,” extending one finger for each count. She then looked her extended fingers and said, “The answer is 9.” c. A student said, “12” and paused. She then counted, “11, 10, 9, 8, 7, 6, 5, 4, 3,” extending one finger for each count. She looked at her extended fingers and said, “They took home 9 slices.” d. A student wrote on her paper 12 – 9 = 3. She said, “I know my fact families, and I know 3 + 9 = 12, so 12 – 3 = 9.”

Feedback from experts on the original item highlighted two issues that were addressed through revision prior to the 2014 field test. First, experts noticed that the original item appeared to be measuring two aspects of MKT: 1) knowledge of the term *counting strategy*, and 2) knowledge of strategies that children who follow the semantic structure of a word problem are likely to use. In order to focus the item squarely on the aspect of MKT most relevant to KCS, reference to the term ‘counting strategy’ in the item stem was eliminated. With the constraint to choose a counting strategy gone, it became necessary to also revise the multiple-choice options, as all but one of the five choices in the original item matched the semantic structure of the word problem. Additionally, one expert reviewer questioned if second grade children would be likely to use the counting down strategy that matched the semantic structure of this word problem when the numbers in the problem necessitate starting at 8 and counting down five counts. He asserted that students are most likely to count down when the number of backward counts is limited to three or fewer. Heeding this observation, we revised the numbers in the word problem to 12 and 3.

Cognitive interview data shed light on additional issues that resulted in revisions. First, respondents struggled to keep track of the information about students presented in the stem of the original problem. Through observing teachers interact with the item, we realized that the information about students was split between the sentences presented before and after the word problem. In efforts to better support respondent comprehension, the stem of the item was revised such that all information about the students was provided in the first sentence, before the word problem. Additionally, cognitive interviews suggested that it took a lot of cognitive effort to comprehend and compare five multiple-choice options. A decision was made to reduce the number of options to four and to make the details of the options as similar as possible on all dimensions except the one of focus (i.e., match to the semantic structure). Ultimately, we decided to focus the first three options on the three types of counting strategies that could be used to solve the problem (i.e., as defined by Carpenter et al., 2015, counting down, counting down to, and counting on to), with each strategy utilizing fingers to keep track of counts.

Item level statistics from the 2014 and 2015 field tests indicated that KCS.PS.2 had good fit to the model. But discrimination values, while within acceptable limits, were borderline low in both field tests. The other items in the ‘predicting student strategies’ category had a similar statistical profile. In efforts to make sense of the low discrimination, we began to consider possible explanations to account for the ‘noise’ in the item. We wondered if teachers’ personal notions of what is taught in first and second grade might be influencing respondents’ ways of interpreting the item. We also began to wonder if these items were really capturing teachers knowledge of content and students or if the focus on match to semantic structure was actually more closely aligned with content knowledge. Ultimately, we have decided to eliminate the ‘predicting student strategies’ category from our scale and are currently attempting to retool the items to eliminate reference to student characteristics and focus more squarely on content knowledge.

Discussion

We set out to create a valid and reliable instrument to measure multiple facets of MKT that we believe important to the teaching of number, operations, and equality concepts in the primary grades mathematics curriculum. Three years into this journey of scale development and refinement, we have learned much about how to improve the quality of items in service of measuring specified constructs. We have found expert review, cognitive interviews, and analysis of field test data to each play a crucial role in our process of improving individual items and the scale as a whole.

Expert review has afforded particularly helpful feedback on the constructs draft items might be measuring and ways to revise items to sharpen the intended focus. Expert review has also informed our thinking about the scale as a whole, as experts have provided feedback on our item categories aligned to various sub-domains of MKT, their relevance to primary grades mathematics instruction, and topics with too much or not enough coverage.

Cognitive interviews have provided us with invaluable insights into how respondent's interacted with our draft items and served as a cornerstone for item and scale development. While expert review improved the underlying substance of items, data from cognitive interviews led to significant improvement in the comprehensibility and formatting of items. Cognitive interviews have also made us much more sensitive to the influence of item context (e.g., the invocation of a particular grade level) on a respondent's way of thinking about a given item. In general, one of the key things we have learned is the importance of clarity in defining what we are trying to measure and removing any unnecessary aspect from the items that might get in the way of the teacher showing us what s/he knows with respect to that topic.

Furthermore, the analysis of cognitive interview data provided valuable insight into how respondents interacted affectively with items. We learned about which items made the respondents groan because of their uncertainty regarding the correct answer. At the same time, there were other items for which respondents provided incorrect answers with much confidence, believing their incorrect answers to be correct. We learned that respondents found enjoyment in items presented with videos of students, and they found visual presentation of student work to be interesting. Cognitive interviews also helped us to gauge the relative amounts of effort and time needed for the various items. All of these insights were helpful in constructing the scale to be used in each field test. We were able to construct a scale for which we had confidence respondents could complete in the targeted amount of time. The data on respondents' affect related to items also informed the ordering of items on the scale. We aimed for respondents to experience the scale such that items identified as having high cognitive demands or low enjoyment were presented in waves, with more enjoyable and easier items in between.

Analysis of field test data using the Rasch model IRT statistics afforded yet another perspective on individual items, item categories, and the scale as whole. Rasch statistics offered insight into the difficulty of items and their fit to the scale as a whole. Items with poor infit/outfit statistics were dropped from the final scale and considered for revision in subsequent field tests. Low discrimination values for individual items prompted renewed scrutiny to better understand how respondents might be interpreting and interacting with the items. In the case of a subset of items in the 'predicting strategies' category, we observed low discrimination values across the set of item; and this prompted a rethinking of what was being measured by those items and ultimately led to dropping the category as originally conceived.

Additionally, our development process has yielded surprising findings related to some of our initial assumptions about inclusion of constructed response items. The 2014 K-TEEM included 7 constructed response items, with one or more items designed for the following categories: 1) Solve Many Ways category in the CCK sub-domain, 2) Interpreting Student Strategies in the SCK sub-domain, and 3) Predicting Student Strategies in the KCS sub-domain. We expected these items to add significantly to the information gained from the multiple-choice and fill-in-the-blank items. Even though we knew these items would require an investment of resources to hand score, we anticipated the benefit to be worth the cost. Instead, we have found our constructed response items to be problematic in multiple ways. Following the 2014 field test, 3 of these items were dropped from the analytic sample. A 'Predicting Student Strategies' item could not be scored reliably, an 'Interpreting Student Strategies' item (the Melody item described in results) had poor item statistics, and one 'Solve Many Ways' item did not appear to be tapping the intended facet of knowledge. In the case of the remaining Solve Many Ways items, we have become increasingly concerned that some respondents may not be willing to take the time to generate responses that demonstrate their true depth of knowledge. This ‘effort factor’ may also be an issue with other constructed response items. We are currently attempting to develop multiple-choice formats to replace some of these items. One promising strategy that we have stumbled on through our experience with SCK.ISS.2 (the Melody item) is adapting responses provided in a constructed response format to multiple-choice options.

Our development process has also led us to question some of our initial assumptions about measuring particular aspects of MKT. One category of items that we currently have aligned with SCK is ‘Interpreting Student Strategies.’ As we have solicited feedback on our measure from persons with strong knowledge of mathematics but little to no experience teaching children, we have observed that these persons have quite competently unpacked the mathematics underlying students’ non-standard strategies. While the ability to interpret student strategies may be more useful in teaching and something that teachers with strong MKT may be able to do more efficiently than others, it does not appear to be knowledge that is exclusive to teaching. Therefore, we are questioning whether it can really be considered SCK.

As has been highlighted previously, we are also rethinking whether select items in our ‘Predicting Student Strategies’ category (for example, KCS.PS.2) are actually measuring teachers KCS or something else. At this point, we are revising the items to drop focus on predicting what students are likely to do and instead focusing the items on simply choosing a student strategy that matches the semantic structure of the word problem. We are hypothesizing that these items will draw on a type of content knowledge that is particularly useful in teaching, but like the ability to interpret students’ strategies, not exclusive to teaching.

Ultimately, we believe an important next step with this scale will be to model the data using factor analysis. A fundamental assumption in item response theory is the unidimensionality of the underlying construct. The more we work with teachers in cognitive interviews, the more we find out that individual teachers have more knowledge in some of these areas than in other areas. Some are stronger in the mathematical content knowledge, such as their knowledge of properties of operations. Others are less knowledgeable in that domain but more knowledgeable in another domain such as recognizing that some types of word problems are more useful for yielding insight into children’s understanding of place value than other word problems. Our experiences are leading us to believe that factor analytic strategies may be better suited for measuring a broad range of MKT than IRT-based methods. Factor analytic methods have an additional benefit in that they may be better suited to providing empirical evidence to inform efforts to further the field’s understanding of constructs within MKT.

Conclusions

We set out to develop a scale that could meaningfully measure the MKT involved in teaching number, operations, and equality concepts in the primary grades curriculum. Based on a combination of our own experiences, an analysis of curriculum standards, and research literature in mathematics education, we think we have made substantial progress in defining the types of knowledge in this domain that primary grades teachers may have and use. We have a fairly large set of items designed to measure whether teachers have this knowledge, and we subjected these items to a one-on-one cognitive interviews and a large-scale field test. The psychometric statistics resulting from the field tests suggest that the resulting scales pass muster with respect to the general standards for educational and psychological measurement.

While we think we are making substantial progress and contributions to the corpus of research in the domain of MKT, there is still a lot of work to be done. The Rasch model accounts for over 75% of the variance in teachers' responses, and the item separation reliability in our measure is quite high. Then again, the item separation reliability (and analogous Cronbach's alpha) are just below the preferred criterion of .80. We think that important next steps will be to reject the assumption of unidimensionality and use factor analysis to run empirical test of the hypothesized subdomains of MKT. We have a large enough sample to run factor analytic models using the data we have from our 2015 sample, and we are optimistic about the prospect of those empirical tests informing MKT theory.

Acknowledgements

The research and development reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305A120781. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Christopher Wolfe, Lynne Nielsen, and Amanda Tazaz made important contributions with respect to development of the K-TEEM assessment items, determining scoring procedures, and performing data modeling and analysis. We are especially grateful for the teachers who took the time to complete the assessment online or participate in cognitive interviews.

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.

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

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (2015). *Children's Mathematics: Cognitively Guided Instruction* (2^{nd} ed.). Portsmouth: Heinemann.

Carpenter, T. P., Hiebert, J., & Moser, J. M. (1981). Problem structure and first-grade children's initial solution processes for simple addition and subtraction problems. Journal for Research in Mathematics Education, 12, 27-39.

Carpenter, T. P. & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education, 15,* 179 – 202.

Carpenter, T. P., Moser, J. M., & Romberg, A. (1982). *Addition and subtraction: A cognitive perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Gersten, R., Taylor, M. J., Keys, T. D., Rolfhus, E., & Newman-Gonchar, R. (2014). Summary of research on the effectiveness of math professional development approaches. Tallahassee: Regional Educational Laboratory Southeast, Florida State University. Retrieved June 30, 2014 from http://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/REL_2014010.pdf.

Hiebert, J. (1982). The position of the unknown set and children's solutions of verbal arithmetic problems. *Journal for Research in Mathematics Education, 13(5),* 341-349.

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., 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., 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 Publishing.

Linacre, J. M. (2006). Data variance explained by Rasch Measures. Rasch Measurement, 20, 1.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington, D.C.: 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. (1960 [1980]). Probabilistic models for some intelligence and attainment tests.[Reprint, with a Foreword and Afterword by Benjamin D. Wright]. Chicago: University of Chicago Press.

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.

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 Publishing.

Turner, E. E., & Celedon-Pattichis, S. (2011). Mathematical problem solving among Latina/o kindergarteners: An analysis of opportunities to learn. *Journal of Latinos and Education, 10(2),* 146–169.