Journal for Research in Mathematics Education
2008, Vol. 39, No. 4, 372–400

Unpacking Pedagogical Content Knowledge: Conceptualizing and Measuring Teachers' Topic-Specific Knowledge of Students

Heather C. Hill
Harvard Graduate School of Education

Deborah Loewenberg Ball and Stephen G. Schilling
University of Michigan

There is widespread agreement that effective teachers have unique knowledge of students' mathematical ideas and thinking. However, few scholars have focused on conceptualizing this domain, and even fewer have focused on measuring this knowledge. In this article, we describe an effort to conceptualize and develop measures of teachers' combined knowledge of content and students by writing, piloting, and analyzing results from multiple-choice items. Our results suggest partial success in measuring this domain among practicing teachers but also identify key areas around which the field must achieve conceptual and empirical clarity. Although this is ongoing work, we believe that the lessons learned from our efforts shed light on teachers' knowledge in this domain and can inform future attempts to develop measures.

Key words: Assessment; Item-response theory; Pedagogical knowledge; Professional development; Teacher knowledge

In the years following Shulman's seminal 1986 address introducing the notion of pedagogical content knowledge (PCK), most scholars and policymakers have assumed that such knowledge not only exists but also contributes to effective teaching and student learning. Standards documents—including those of NCTM and the National Board for Professional Teaching Standards (NBPTS)—note the importance of teachers holding knowledge of "students as learners" (NCTM, 2000, p. 17)

Measures development supported by NSF grants REC-9979873, REC-0207649, EHR-0233456, and EHR 0335411, and by a subcontract to CPRE on Department of Education (DOE), Office of Educational Research and Improvement (OERI) award #R308A960003. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies. The authors thank three anonymous reviewers and members of the LMT project for reading and critiquing drafts of this article. Patrick Callahan and Rena Dorph also provided generous assistance with this project. Errors are our responsibility.

and being able to recognize the "preconceptions and background knowledge that students typically bring to each subject" (NBPTS, 2006, p. vi). Preservice programs and professional development opportunities often focus on developing teachers' knowledge of and skill in understanding students' mathematical work and thinking.

Yet scholarly evidence about what PCK *is*, and how it relates to students' mathematical outcomes, is actually quite thin. Although well-designed research has shown that professional development focused around such knowledge results in changed classroom performance and improved student learning, these results are limited to a handful of studies in relatively narrow content areas (Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Cobb et al., 1991; Franke, Carpenter, & Levi, 2001; Saxe, Gearhardt, & Nasir, 2001). And beyond Shulman's original formulation, there have been few attempts to develop concise yet well-specified descriptions of what teachers know in this domain. Further, no large-scale study has related teachers' PCK to students' gains. Although what we call "knowledge of content and students," or teachers' knowledge of students' mathematical thinking and learning, is widely believed to be an important component of teacher knowledge, it remains underspecified, and its relationship to student achievement undemonstrated.

We argue that these gaps stem from a two-fold problem in mathematics education research. First, we lack studies that demonstrate that teachers possess this knowledge *apart from knowledge of the content itself*. Second, the field has not developed, validated, and published measures to assess the many programs designed to improve teacher knowledge in this domain and to understand how this knowledge relates to student achievement.

This article chronicles a first effort to conceptualize, develop, and test measures of teachers' knowledge of content and students (KCS). We do so in a framework that ultimately connects all three pieces of this work, tying the conceptualization directly to the specification of items, and tying results from field tests back to strengths and weaknesses of the initial conceptualization. Although this is ongoing work, we chose to write about it at this particular juncture because our efforts might be instructive to others trying to conceptualize, identify, measure, and ultimately improve teachers' PCK. This work might also be useful to those engaged in parallel work, such as measuring teachers' ability to design effective instruction and measuring teachers' skills in motivating students to learn mathematics. Finally, this work is an important precursor to designing and implementing large-scale studies that assess whether teachers' knowledge of mathematics and students contributes to student learning. As economists and others have noted, teacher characteristics such as credentials, experience, and even general mathematical knowledge provide only limited explanations for wide variations in student gain scores across classrooms (for a review of the economics literature, see Wayne & Youngs, 2003; for studies specifically focused on mathematics, see Begle, 1972, 1979; Monk, 1994). Educators, and mathematics educators in particular, need to do more to help explain this phenomenon—beginning with developing new, more sensitive instruments that capture key teacher characteristics.

We argue that two sets of criteria are vital to building such measures. The first set of criteria concerns the conceptualization of the domain. Researchers should begin by proposing a construct, taking care to elaborate the theoretical or empirical basis for the construct, delineate the boundaries of the construct, and specify how it is related to other similar constructs. Each of the above should provide relatively specific information regarding the nature of knowledge and how it might be assessed. As part of this work, researchers should consider the cognitive process involved in this domain, particularly with regard to how measurement should proceed. Measures of facts, for instance, will look different from measures of reasoning processes; measures of knowledge that is highly contextually bound will look quite different from measures of knowledge that is common across a wide range of settings.

Second, in any measures development effort, data yielded from pilots of the items must be analyzed to assess whether the conceptualization is correct and adequate and to determine whether the instruments meet several measurement-related criteria. These criteria include that construct-identification methods (e.g., factor analysis) demonstrate that the construct is clearly present and separable from other related constructs; that the items adequately measure the population under study, in the sense that they provide a reliable estimate of individual knowledge across a range of levels of expertise; and that validation work begin to assess, using nonpsychometric methods, whether the items tap the intended construct.

These criteria are based on those found in the AERA/APA/NCME *Standards for Educational and Psychological Testing* (1999) and informed by current debates in the field of educational measurement and evaluation (e.g., Kane, 2004). We believe that meeting these criteria is critical to yielding credible measures. Yet this kind of effort is seldom launched (Messick, 1988). All too often, scholars report results from locally developed measures without reference to their conceptualization, design, psychometric, and validation work. We also argue that fulfilling these criteria can yield important information about the constructs themselves and the populations under study.

This article follows an unusual format. We begin by discussing our conceptualization of the domain of KCS; we regard this conceptualization as a hypothesis to be explored in the empirical portions of the article. We then describe our efforts to write items and outline the methods used in piloting, analyzing, and validating these items. In the third section, we assess whether our conceptualization is accurate, asking whether KCS exists in the general population of teachers and, if so, how we might describe it. We also ask whether the measures have sufficient validity and reliability for use in research studies.

CONCEPTUALIZING THE DOMAIN

Our project seeks to understand and measure mathematical knowledge for teaching—the mathematical knowledge that teachers use in classrooms to produce instruction and student growth. Our work to date has focused mostly on teachers'

subject matter knowledge—not only knowledge of the actual topics they teach but the special forms of mathematical knowledge that are particular to the profession of teaching (Ball, Hill, & Bass, 2005; Hill, Rowan, & Ball, 2005). In addition to subject matter knowledge, however, we believe that teachers might also possess additional forms of knowledge useful to their work in classrooms. Below, we describe how we conceptualized one such strand of this knowledge, taking care to discuss its theoretical and empirical justification, to delineate its boundaries and relationships to other constructs, and to discuss the measurement implications of our conceptualization.

Defining KCS

We propose to define KCS as content knowledge intertwined with knowledge of how students think about, know, or learn this particular content. KCS is used in tasks of teaching that involve attending to both the specific content and something particular about learners, for instance, how students typically learn to add fractions and the mistakes or misconceptions that commonly arise during this process. In teaching students to add fractions, a teacher might be aware that students, who often have difficulty with the multiplicative nature of fractions, are likely to add the numerators and denominators of two fractions. Such knowledge might help her design instruction to address this likely issue. In thinking about how students might solve a problem like 56+9, to use another example, a teacher might know that some students will count on, some will add 10 and then compensate by subtracting 1, and still others will use a standard algorithm.

This definition is based on both theoretical and empirical work on teacher knowledge. To start, KCS is a primary element in Shulman's (1986) PCK. In this view, such knowledge is composed of "an understanding of what makes the learning of specific topics easy or difficult: the conceptions and preconceptions that students of different ages and backgrounds bring with them to the learning of those most frequently taught topics and lessons" (1986, p. 9). Shulman noted that research on students' thinking and ideas provides a critical foundation for pedagogical knowledge.

In mathematics, the work of Fuson (1992), Kamii (1985), and Carpenter, Fennema, Franke, and Empson (1999) in the areas of number and operation, Behr, Harel, Post, and Lesh (1992), and Lamon (1999) in rational numbers, and Carpenter, Franke, and Levi (2003) in algebra exemplify research on how students solve problems, develop mathematically, and encounter difficulties with particular aspects of subject matter. Following Shulman, we used these and other content-specific studies of student learning as the foundation for the measures described below. We chose *not* to ground our measure in overarching and thus more generic theories of learning (e.g., constructivism or behaviorism) for two reasons. First, our definition of KCS suggests that we rely on *empirical* evidence regarding how students learn. Theory is, literally, theory, and is thus propositional and arguable; "correct" answers based on theory would be difficult to defend, for several theories of student learning

legitimately compete. Teachers' answers would be indicative of views of student learning, or knowledge of such views. Second, theories of student learning are also necessarily abstracted from specific instances, making them difficult to use in item-writing. By contrast, we wanted to measure teachers' knowledge with demonstrated developmental patterns and problems with specific material.

Evidence for KCS

The strongest empirical evidence for KCS comes from experimental and other tightly controlled professional development studies in which teachers investigated how students learn particular subject matter, such as whole number operations or fractions (Carpenter et al., 1989; Cobb et al., 1991; Franke & Kazemi, 2001; Saxe et al., 2001). When teachers studied this material, their classroom practices changed and student learning was improved over that of teachers in control or comparison groups. This suggests that such knowledge is useful to teachers' disciplinary teaching. However, it says nothing about whether teachers who are *not* involved in such professional development possess such knowledge and, if so, what shape it takes. It is also possible, in many of these programs, that teachers learned some mathematical content itself. In this case, improvements in subject matter knowledge, rather than KCS, would be driving student achievement.

In fact, there are only a limited number of investigations into what "average" teachers know about students' mathematical thinking. A search of "pedagogical content knowledge" in the Educational Abstracts database led to only one study directly on this point. Carpenter, Fennema, Peterson, and Carey (1988) explored 40 first-grade teachers' knowledge of children's solutions of addition and subtraction word problems. Their result was slightly paradoxical: Although most teachers could distinguish between problem types and predict which would be relatively more difficult for students to solve, much of this knowledge was tacit. For instance, teachers had difficulty articulating *why* specific problems would be difficult for students. The authors concluded that participants' knowledge was not organized into a coherent network connecting the mathematics problems to student solution strategies.

Two other studies explored preservice, rather than in-service, teachers' understandings of student learning. In a study of division of fractions, Tirosh (2000) found that prospective elementary teachers in Israel were familiar with common arithmetic bugs (e.g., inverting the dividend rather than the divisor) but unfamiliar with conceptual errors made by students, including the overgeneralization of whole number rules to fractions (e.g., dividing always makes numbers smaller). In a study that asked preservice teachers to evaluate the difficulty of algebra problems, Nathan and Petrosino (2003) argued that teachers with advanced subject-matter knowledge of algebra were likely to believe that students find word problems more difficult than symbolic problems. This result contradicts research on actual student learning conducted by Nathan and others (e.g., Koedinger & Nathan, 2004). The thinness of teacher knowledge revealed in these studies may result from the

population under study, however, and not reflect the true knowledge level of practicing teachers.

These studies suggest intriguing hypotheses—for instance, that teachers' knowledge of students is propositional and discrete rather than richly connected to the underlying mathematics and conceptions of student learning. However, none of these studies explicate the domain of teacher knowledge as carefully as necessary for large-scale measurement and modeling vis-à-vis student achievement, as many hope to eventually accomplish. This interest led to our own efforts to conceptualize KCS; the content of these studies, however, helped shape our definition.

Relating KCS to Other Forms of Teacher Knowledge

Our criteria suggest that an important element in conceptualizing a domain is delineating its boundaries and relating it to similar constructs. Figure 1 shows our proposed model of mathematical knowledge for teaching (MKT) and can be used to demonstrate how KCS relates to both subject matter knowledge and PCK. Each of the six portions of the oval is a proposed strand of MKT. The left side of the oval, labeled "subject matter knowledge," contains two strands that lie *outside* Shulman's popular conceptualization of PCK: *common content knowledge* (CCK), roughly described as knowledge that is used in the work of teaching in ways *in common with* how it is used in many other professions or occupations that also use mathematics, and *specialized content knowledge* (SCK), or the mathematical knowledge that allows teachers to engage in particular *teaching* tasks, including how to accurately

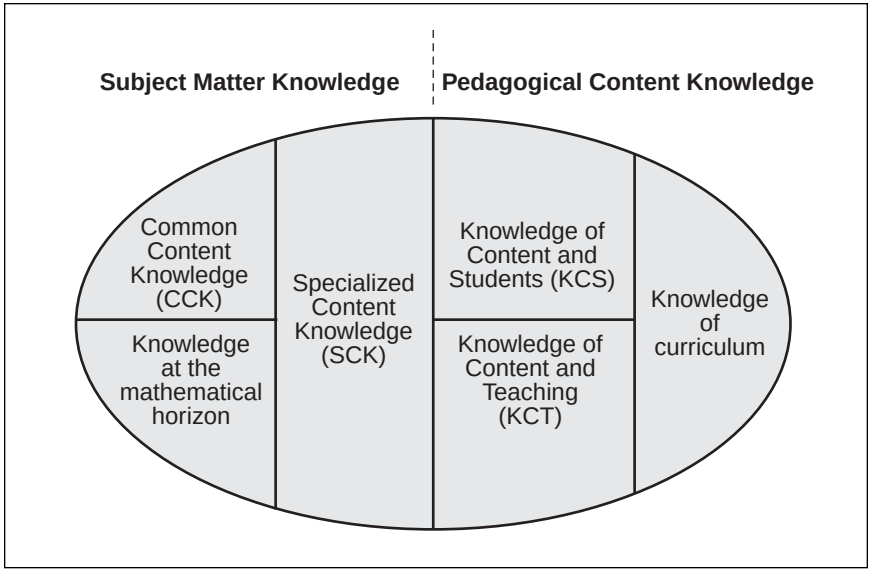

Figure 1. Domain map for mathematical knowledge for teaching.

represent mathematical ideas, provide mathematical explanations for common rules and procedures, and examine and understand unusual solution methods to problems (Ball et al., 2005). CCK is what Shulman likely meant by his original *subject matter knowledge*; SCK is a newer conceptualization. However, both are mathematical knowledge; no knowledge of students or teaching is entailed. The right side of the oval represents strands associated with Shulman's proposed PCK and contains KCS, knowledge of content and teaching (KCT), and knowledge of curriculum. KCS is thus a subset of PCK, which itself is a subset of the larger construct MKT. KCS, however, is separable from knowledge of teaching moves—for example, how best to build on student mathematical thinking or how to remedy student errors. KCS also does not entail knowledge of curriculum materials. Instead, it is focused on teachers' understanding of how students learn particular content.

Thus, a key part of our conceptualization holds that KCS is distinct from teachers' subject matter knowledge. A teacher might have strong knowledge of the content itself but weak knowledge of how students learn the content or vice versa. This distinction has important implications for the construction of assessments in this domain, as described next.

Developing Criteria for Developing Measures

Having outlined the theoretical and empirical basis for KCS and discussed its relationship to similar constructs, we turn now to developing the notion further for the purpose of measurement. In keeping with the conceptual distinction between KCS and subject matter knowledge, we defined the KCS domain by stipulating that in order to analyze an item and arrive at their answer, respondents should use knowledge of students' thinking around particular mathematics topics, rather than purely their own mathematical knowledge, test-taking skills, or other processes. Because KCS is an amalgam of subject matter knowledge and knowledge of students, we expect that teachers might also invoke mathematical knowledge or engage in mathematical reasoning in order to interpret students' thinking around these topics. However, they should not *solely* engage in mathematical reasoning when answering these items—they must also invoke knowledge of students.

Further developing KCS for measurement opened several debates as to what should be in the domain. One debate concerns whether the established literature on student learning should be the *only* basis for items. It seems reasonable that teachers might have KCS not yet discovered by academics, and, in keeping with this theory, we included the knowledge that observant teachers might glean from working with students, but that have not been codified in the literature. On the other hand, such knowledge has by definition not been codified, making the process of actually writing items based on this knowledge hazardous. How would we know that what item-writers had observed in classrooms held true generally? The result was only a small number of such items.

A second debate concerns whether to measure teachers' recognition or recall of topics in this domain, such as "knowing that" students tend to develop in a certain

way or make particular mistakes or whether to measure teachers' ability to *reason* about student work. Because we were working from a view that KCS stems in large part from the educational psychology literature, we erred on the side of items that tap established knowledge. As we will report below, cognitive interviews with teachers raised questions about this assumption.

Throughout the early conceptualization of these items, we struggled with how much to contextualize the KCS items. If Shulman and others are correct in thinking that educational psychologists have uncovered regularities in student mathematical learning and errors, these regularities should appear independently of the teaching methods or curriculum materials used. In addition, we quickly saw that adding even a modest context (e.g., naming specific curriculum materials or describing previous lessons) not only made the items intolerably long but also might disadvantage teachers not familiar with particular materials or techniques. In our pilot, then, KCS was centered on the development and errors made by the modal U.S. student, regardless of the curriculum materials, instructional techniques, and other mathematical influences the student might have encountered. This contextual independence is a key, and perhaps problematic, feature of our conceptualization of knowledge in this domain.

Our conceptualization of KCS will, of course, differ from others'. However, we argue that by being explicit in this formulation, we provide both a foundation for our measures development work and a basis for future discussions about the nature of teacher knowledge. We turn now to how we operationalized this conceptualization into items.

WRITING ITEMS AND GATHERING DATA TO TEST OUR CONCEPTUALIZATION

An important phase in our measures development project was moving the conceptualization of the domain into practice—writing items and testing them with large groups of teachers. With data in hand, we then needed to apply our measurement criteria, searching for confirmation that we were, indeed, measuring a real construct distinct from teachers' general subject matter or pedagogical knowledge and that our measures met basic psychometric criteria. In this section, we discuss the process of writing items, determining what evidence to gather to evaluate the measurement criteria, and collecting and analyzing data.

Translating the Conceptualization Into Items

Once the larger domain map had been set, we began writing multiple-choice items based on the literature cited previously and on our own classroom experiences. We chose the multiple-choice format because one of the projects for which we were designing items included over 5,000 teachers, many of whom responded to the items multiple times over a period of 3 years. In this context, open-response items, which entail significant "grading" of answers, were not feasible. In developing items, we

found that it was helpful to think about what mathematically able individuals who do *not* teach children would not know. For instance, although professional mathematicians would know how to produce a definition of even numbers, multiply a two-digit number by a one-digit number, and write five hundred and twenty-six as 526, they may not know the grade levels at which students, on average, master these ideas and tasks. They are also unlikely to be familiar with common errors that students make while developing proficiency with these ideas and procedures.

As we progressed, we saw that our items tended to fall into four major categories:

- Common student errors: identifying and providing explanations for errors, having a sense for what errors arise with what content, etc.
- Students' understanding of content: interpreting student productions as sufficient to show understanding, deciding which student productions indicate better understanding, etc.
- Student developmental sequences: identifying the problem types, topics, or mathematical activities that are easier/more difficult at particular ages, knowing what students typically learn "first," having a sense for what third graders might be able to do, etc.
- Common student computational strategies: being familiar with landmark numbers, fact families, etc.

The Appendix contains four sample items. These items are imperfect; statistical analyses showed they were not sufficiently related to the construct measured by the majority of piloted items to retain in our item pool. The reader may be able to see why these items failed just by reading them. They are presented here, however, to give a sense for how we represented the four major categories in item-writing.

In item 1, teachers are asked to provide the most reasonable explanation for Bonny's error. Although Bonny can count out 23 checkers serially, she does not understand that the "2" in 23 indicates two groups of ten, or 20 checkers. Evidence suggests that although Bonny understands how large 23 is (she can count and represent 23 serially), she does not understand the meaning of the places in the number 23. This misunderstanding has been documented as common to children in the early elementary grades (Kamii, 1985, p. 61) and may be quite familiar to teachers experienced with working with students in this particular area.

In item 2, Mrs. Jackson must analyze three students' work on multidigit addition problems to determine who has made the same error. Students (I) and (II) have made the same error, carrying a 10 instead of a 20; student (III) has failed to add numbers in the ones column correctly. Both errors are common, according to the "buggy algorithm" literature and teachers interviewed for this project; as well, diagnosing student errors in computation is a common task of teaching.

In item 3, Mr. Fitzgerald intends to design an assignment that would show him whether students can correctly order decimals. We intended here that Mr. Fitzgerald would recognize that students will be able to put options (a) and (b) in order correctly while ignoring the decimal point. Such students, according to Resnick and

colleagues (1989), are applying whole number rules to decimals. Only option (c) would require students to understand, for instance, that .565 is a smaller number than 4.25 to order the decimals correctly.

Finally, item 4 is an open-ended item that asks teachers to explain Jill's incorrect answer to the subtraction problem 51 – 18. This item was not piloted as part of our large-scale measures development but instead included as part of our cognitive interviews (see below), both to glean information on how the format of an item might affect its performance and to develop response choices for the item. To get this item correct, teachers would have to recognize that Jill "subtracted up" (8 – 1 = 7), rather than trading tens for ones. This error has been identified as a common "buggy algorithm" by the cognitive psychology literature on this topic (Ashlock, 2002; VanLehn, 1989).

Collecting and Using Data to Test Our Conceptualization

At the outset of this article, we argued that any measurement effort must collect and use data to test the conceptualization of the measures, to determine whether the measures do in fact differentiate well among individuals, and to perform validation work. In our case, this effort took several directions. At a minimum, we felt we needed to collect data from large samples of teachers to conduct construct identification procedures and reliability analyses. Because we collected this data as part of an evaluation, it also allowed us to complete a related validity check: determining whether teachers' growth on our measures is sensitive to their reports of learning about how students learn mathematics. Later, we added cognitive interviews to provide checks on our assumption that teachers use knowledge of students to answer the items.

Construct identification and scaling. We piloted the items in California's Mathematics Professional Development Institutes (MPDIs), where our larger set of measures (both around subject matter knowledge and KCS) served as part of the evaluation of that program. Initiated in 2000, the MPDIs involved both mathematicians and mathematics educators in the design and implementation of content-focused, extended learning opportunities for teachers. Teachers attended summer programs of 1 to 3 weeks' duration (between 40 and 120 hours), participated in up to 80 hours of school-year follow-up, and received stipends of approximately $1500 for their participation.

Although a focus on student learning of mathematics was only a tertiary goal of the MPDIs, where the main focus was on improving teachers' subject matter knowledge, visits to several MPDIs over the course of this investigation revealed that some had, in fact, included a focus on student learning of content as part of their curriculum. In one site visited, for instance, teachers examined student computational work for errors, explained those errors, and discussed how they would remedy them in teaching. In another, teachers studied the literature describing which problems students found difficult and easy and examined methods that students might use to solve various types of problems. In still other sites, teachers viewed

videotape of students solving problems and examined student work. However, work on student thinking and learning was not universal. In several other sites observed, no focus on student learning was apparent. This variation is actually to our benefit, as we can use it to examine the impact of different amounts of opportunity to learn these topics on teachers' KCS performance.

By legislative design, every teacher attending an elementary number and operations MPDI was to have completed a pre-post evaluation form. These forms were designed by our project and contained items intended to assess both teachers' subject matter knowledge and KCS in elementary number and operations. Teachers cycled through parallel forms of the assessment during the summer portion of the program, with one group taking form A as a pretest and B as a posttest, another group taking B as a pretest and C as a posttest, and so forth. The parallel form structure mitigated against memory effects, where teachers retaking the same form might recall their answers to the previous form or have discussed answers with others prior to the posttest.

Numerous problems plagued the 2001 administration of the pre-post assessments, including missing booklets and teacher and professional development provider refusals (see Hill & Ball, 2004, for details). Nevertheless, by combining all reasonably complete pretests and posttests, we reached 640 responses to form A, 535 responses to form B, and 377 responses to form C. Each form contained items in (a) number and operations common and specialized content knowledge; (b) number and operation KCS; and (c) patterns, functions, and algebra common content knowledge. By including multiple hypothesized domains on the assessment instrument, we were able to conduct the construct identification analyses described below. We also included a set of "teacher opportunity to learn" items in which teachers reported on the content of their MPDI. These reports were used in the convergent and discriminant validity analysis, described below.

The first analysis conducted with this data was scaling work, including factor analysis and item response theory (IRT) measure construction. These analyses allowed insight into a central question of this article: Do items capture KCS or do these items measure a more general dimension, such as overall mathematical knowledge? And by extension, does the general population of teachers have a kind of "knowledge" that can be labeled KCS? Factor analysis can help identify constructs, first determining the number of separate constructs on forms, then providing information on which items relate to which constructs. On our forms, for instance, there might have been only one construct despite the presence of both CK and KCS items; if so, we would have said that there is no distinction in our data between these two theoretical categories. Or a factor analysis might indicate the form contains two constructs, and items may group onto "CK" and "KCS" factors, as theorized. Other possibilities exist, and we used ORDFAC (a variant of TESTFACT, a common factor analysis program for test data; Schilling, 2005) to help us sort among these possibilities.

Once factor analysis confirmed the number of constructs on the form, we proceeded to scale items. Our factor analysis did indicate multidimensionality,

meaning scaling ideally would proceed using multidimensional IRT models. However, such models require more than a thousand cases for proper estimation, data we did not have at the time and were reluctant to collect, given the fact that this was the first pilot of such items. Instead, we used Bilog (Zimowski, Muraki, Mislevey, & Bock, 2003), software commonly used in unidimensional scale construction. Bilog reports two useful pieces of information for understanding how well a set of items measure a construct in a given population. One is the reliability of the items, or how accurately the items distinguish between individuals. Reliability is reported on a 0–1 scale, with lower values indicating scales that either cannot distinguish accurately between differences in individual knowledge, or can only accurately distinguish gross differences in knowledge (e.g., between a scholar of student learning and a novice teacher). Higher values indicate an ability to more precisely distinguish between individuals closer together on the underlying construct (e.g., two teachers with only subtle differences in their level of knowledge). The second useful descriptor is the test information curve maximum. If one imagines a hypothetical normal distribution of teachers along an *x*-axis from lowest-knowledge to highest-knowledge, the test information curve maximum corresponds to the point on that distribution where the items' ability to differentiate among teachers is the highest. Ideally, a test information curve for a general-purpose assessment would be centered on the peak of the normal distribution, since that is where the majority of individuals' knowledge level is located. However, assessments may be more difficult—with their test information curve peaking at a standard deviation or more above the average teacher—or less difficult. This information is critical in examining how useful the assessment will be in research and evaluation.

Convergent and discriminant validity. We also used data from the MPDIs to conduct convergent and discriminant validity checks. In convergent validity, analysts seek to determine whether constructs that should relate to one another actually do so; in discriminant validity, analyses confirm that constructs that should not be related are, in fact, not. In our case, any gains in teachers' KCS performance should relate to their reports of having had opportunities to learn KCS but not to learning other things, such as reports of having had opportunities to learn subject matter knowledge. To determine whether this was the case, the same instrument that carried the KCS items also asked teachers to describe their MPDI's content. Two Likert items asked teachers whether they focused on how students learn number concepts and operations; these formed a scale with reliability .73 and mean 3.24 on a 1 (low) to 4 (high) scale, indicating that many teachers reported focusing on these topics. Another two items asked teachers to report on how much their MPDI focused on subject matter knowledge in number and operations; these items had a reliability of .60 and a mean of 3.44, suggesting that this was also a major focus of the workshops. A single item asked teachers to report the extent to which the MPDI contained "student work for participants to study"; the mean of this item is 2.42 on a scale of 1 (low) to 4 (high). Finally, four items asked teachers the extent

to which the MPDI provided an opportunity to learn purely mathematical topics, including proof, justification, representation, and communication. Although this variable positively predicted growth in teachers' subject matter knowledge during the institutes (Hill & Ball, 2004), it should not predict teacher growth in KCS. Here, its mean is 3.28 on a scale of 1 (little opportunity) to 4 (much opportunity) and its reliability is .74. Finally, information about the length of institutes was added to the data from administrative records.

To enable this validity analysis, our parallel forms were equated using conventional IRT linking methods (see Hambleton, Swaminathan, & Rogers, 1991; McDonald, 1999). Details on the KCS scales in number and operations are reported in Table 1. Teachers' ability is reported in standard deviation units, with 0 representing the average teacher, +1 representing a teacher one standard deviation above average in KCS, and so forth. To explore reasons for differential growth in teacher knowledge, we ran linear mixed model regressions (Bryk & Raudenbush, 1988; Singer, 1998) using teachers' performance on the posttest as an outcome variable, and predicted these outcomes using pretest score and the MPDI content indicators described earlier. These models were run in SAS (SAS Institute Inc., 1999).

Cognitive interviews. In addition to determining convergent validity, we were also interested in a much more rudimentary form of validity: determining whether what we conceptualized as KCS was actually measured by the items. Instead of tapping individuals' KCS, for instance, our items might measure individuals' capacity to use test-taking strategies. To determine whether such alternative hypotheses were true and to learn more about the knowledge used to answer the items, we turned to cognitive interviews.

A full account of sample recruiting, characteristics, instrumentation, and procedures can be found in Hill, Dean, and Goffney (2007); we briefly describe each here to provide context for the results below. Fifty-seven K–6 teachers were recruited from three Midwestern school districts—two serving disadvantaged student populations and one serving a socioeconomically mixed group of students. Fifty of these teachers returned survey forms; from these 50, we selected 26 for interviews on the basis of either low or high CK scores. Interviews took place from 1 week to 2 months after the return of the surveys. During the interviews, teachers were asked to report on how they answered questions related to 17 stems; the exact probe was "Why did you choose [answer choice]? What process did you go through to decide?" This protocol constitutes what Sudman, Bradburn, and Schwartz (1996) would term a "retrospective think-aloud," and provides a rough sense for the major characteristics of the cognitive processes respondents engaged as they answered items.

Subsequent to transcribing and entering teachers' interviews into QSR N6 (QSR International, 2002), a program that facilitates the management, coding, and analysis of qualitative data, we categorized each teacher's response to each question based on the knowledge invoked to answer. Coding for each item was done by categorizing teacher responses until coders reached 70% or greater agreement, at which

point one individual completed coding the rest of the item. This process allowed coders to come to agreement on how the coding scheme should be applied to particular types of answers within each item. The codes themselves, described below, were developed by reading, discussing and coding teachers' responses to a variety of items.

RESULTS: APPLYING THE MEASUREMENT CRITERIA

Identification of a KCS Construct

We first asked whether we could identify a KCS construct in the data. This is an important question: If the items intended to measure KCS appeared no different from the items intended to measure subject matter knowledge, we could not claim that they measured the body of knowledge as we conceptualized it. Conversely, finding a KCS factor would suggest that teachers do have discipline-specific knowledge of student learning, as many have hypothesized.

Extensive results on factor analyses are included in Hill, Schilling, and Ball (2004) and Schilling (2007) and will be only briefly reviewed here. Exploratory factor analyses *did* indicate that the KCS items formed their own separate, interpretable factor but with a wrinkle: Some items meant to tap teachers' KCS scaled with items meant to tap CK, although no obvious differences emerged between these items and those that scaled on the KCS factor. In a specialized confirmatory factor analysis, where items were allowed to load either on a "general" or "KCS" factor, most items loaded on both, suggesting either that (a) some individuals used KCS and others used subject matter knowledge to answer these items, or (b) individuals used a combination of both KCS and subject matter knowledge to answer the items. This indicates that at least in the population that responded to our instruments, teachers used both subject matter knowledge and KCS to answer these items. It also suggests an answer to one of the central questions of this article: There *is* an identifiable "knowledge of content and students" within the teaching population, at least with the particular set of items piloted and in the particular population of teachers included in our sample (Hill et al., 2004). However, this knowledge may rely in part on teachers' underlying subject matter knowledge and is imperfectly discerned with the set of items used in our current instruments.

Meeting Standards for Reliability

The next step in analysis was to determine the reliability of the KCS items on each form. Reliability is a critical diagnostic in assessing the utility of a set of measures, and it can be thought of intuitively in several ways. Most formally, reliability is defined as the proportion of true score variation in the data to true score and error variation combined, or the ratio of "signal" to "signal + noise" in a measure. Noise may result from several sources: (1) a set of items that do not cohere as well as intended (e.g., a set where 10 items measure construct A and another 5 measure construct B); (2) the presence of items that fail to help distinguish between

knowledgeable and less knowledgeable individuals; (3) a set of items that are mismatched to the ability level of the population being measured. In general, reliabilities of .70 or above are considered adequate for instruments intended to answer research and evaluation questions using relatively large samples.

Reliabilities for the forms were low, relative to industry standards. In a previous article, we reported IRT reliabilities as .71 for form A, .73 for form B, and .78 for form C (Hill et al., 2004). These reliabilities described a sample that combined pretests and posttests for each form, a decision we made in order to allow the use of an IRT model that overweights highly discriminating items in the estimation of reliability. Table 1 shows the reliabilities for teachers and forms included in the pre-post evaluation reported below. Because we had a smaller sample size,[1] the IRT model chosen did not overweight highly performing items, and reliabilities were slightly lower than our original report. These reliabilities were lower than industry standards suggest are sufficient for use in research and evaluation projects. Said another way, these measures cannot as accurately discriminate among teachers of different knowledge levels as we would like, often misplacing teachers relative to one another and to specific benchmarks (e.g., an average teacher).

Table 1
Reliabilities of KCS Scales in Number and Operations[a]

Scale	Number stems	Number items	Reliability	Test information curve max[b]
Form A				
Pretest	14	20	.60	–1.5
Posttest	14	20	.67	–1.12
Form B				
Pretest	14	19	.68	–1.0
Posttest	14	19	.65	–1.25
Form C				
Pretest	15	21	.58	–1.37
Posttest	15	21	.69	–0.875

[a]These are one-parameter IRT models. Some items are nested beneath one common stem, e.g., Appendix item 1. Thus the number of items is more than the number of stems on each form.
[b]These test information curve maxima are reported prior to form equating.

One possible reason for these lower reliabilities is the multidimensionality found in the factor analysis. When a set of items draw upon more than one type of knowledge (i.e., CK and KCS), the items do not cohere as tightly as if all measured the same construct. But other explanations are also suggested by diag-

[1] "Two-parameter models," in which items that strongly discriminate between teachers are overweighted, typically require several hundred respondents for proper estimation. Here we used one-parameter models, which do not weight items. When each item is weighted equally, reliabilities fall because the proportion of the scale composed of poorly performing items is increased.

nostic information about the overall scales and individual items. One common practice in IRT analyses is to examine each item's ability to discriminate between teachers at a particular ability level using a metric called the "slope" (for a technical interpretation, see Hambleton et al., 1991). Generally, items must have slopes above .4 for continued use. Across all three forms, roughly 15% of items failed this criterion. This suggests that our items often failed at capturing elements of knowledge about students that existed in the population of teachers participating in this study; in other words, much of what we hypothesized to be an aspect of this knowledge did not help distinguish between knowledgeable and less knowledgeable teachers in this domain.

Another explanation exists for the relatively low reliabilities. The test information curve, which identifies the level of knowledge at which the scale best measures individuals, suggests that for all three forms, best measurement occurs for teachers between 1 and 2 standard deviations below average. Figure 2 shows this in more detail. The *x*-axis is the teachers' scale score; 0 typically corresponds to the average teacher in the population under study, with negative scores indicating less knowledgeable teachers and positive scores indicating more knowledgeable teachers. In this case (form A pretest), the most information (shown as a solid line and measured on the left *y*-axis) is provided for teachers between two standard deviations below average; a corresponding standard error estimate (shown as a dotted line and

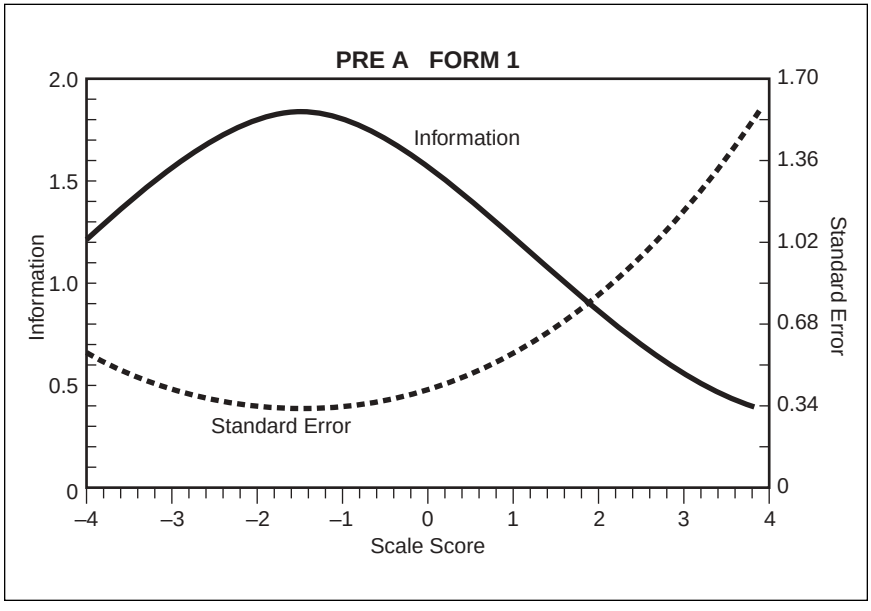

Figure 2. Test information curve.

measured on the right *y*-axis) shows that errors in score estimation rise precipitously for individuals above average.

This finding has practical implications. First, identifying highly knowledgeable teachers and tracing growth among groups of teachers who become highly knowledgeable will be difficult. A second implication relates back to how item construction took place. Because we hoped to be able to make distinctions among teachers of all knowledge levels, we wrote items intending that the average score on the assessment would be 50%. This design implied including difficult items, ones that perhaps only 20–40% of the population would answer successfully, in order to discriminate between highly knowledgeable teachers and very highly knowledgeable teachers. Instead, on only 13 of 59 items did teachers answer correctly less than 50% of the time—and only 3 items had a frequency correct of 25% or fewer. This is quite off the target, especially considering that many of the items teachers "missed" were ones that our analyses showed did not actually capture teachers' KCS. Reasons for this are explored in more detail below.

Capturing KCS as Conceptualized

Factor analysis and measure construction provide one method for checking the validity of these items—that is, whether the hypothesized ability we measured is actually a coherent domain. If results accord roughly with theory, the theory is neither disconfirmed nor confirmed; the factors we identified, for instance, might represent test-taking skills or even reading ability. To learn more about what we measured, our cognitive interviews contained six KCS items focused around common student errors in number and operations; five were multiple choice and one was open-ended (see Appendix item 4). Hill et al. (2007) described the complete coding scheme used to classify teacher answers; we include here only the explanations prevalent in answering KCS items:

- KCS—Teacher invokes familiarity with students' errors as partial or complete explanation for selecting their answer (e.g., "My students do this all the time.")
- Mathematical reasoning—Teacher uses mathematical deduction, inference, or other type of reasoning to support her answer (e.g., "Looking at these numbers listed in the problem, (c) must be true.")
- Test-taking skills—Teacher uses information in the stem, matches response choices to information in the stem, or works to eliminate answers as method for solving problem (e.g., "I knew it wasn't (a), thus (d) 'all of the above' couldn't be true.")

Answers could be coded more than once—for instance, if a teacher used both test-taking skills and KCS. Additionally, answers coded "KCS" had to be the KCS the item was meant to capture rather than knowledge of students used to support an "incorrect" answer, although subsequent analyses relaxed this assumption.

Table 2 shows the percentage of teacher responses for each item. Clearly, KCS was an often-supplied explanation for answers; in 42% of cases, teachers referenced

their experience with children to support their chosen answer. In some cases, the reference was spontaneous. While discussing Jill's buggy subtraction (Appendix item 4), a teacher said, "She's just subtracting the smaller digit in each place from the larger, not regrouping. And I see a lot of that." In others, an interviewer probe triggered teachers' statement of familiarity with the student error (e.g., "*I:* And that's a common problem you've seen with your kids?" *R:* "Huge. Huge. Huge."). Not surprisingly, teachers' familiarity with these student errors was greatest when they taught children of similar age and developmental levels. This suggests that teachers do have knowledge of students' common errors and that these items partially tapped this knowledge.

Table 2
Reasons for Item Responses—Validation Interviews

Items	KCS	Mathematical reasoning	Test-taking skills	Row total
Bonny's problem with 23	13 (37%)	9 (26%)	13 (37%)	35
Student error with base-10 blocks	18 (55%)	6 (18%)	9 (27%)	33
Mrs. Jackson always carry 1	16 (38%)	25 (60%)	1 (2%)	42
Student misunderstanding of equal sign	16 (52%)	6 (19%)	9 (29%)	31
Ordering decimals assignment	1 (4%)	25 (93%)	1 (4%)	27
Jill's buggy subtraction	19 (61%)	12 (39%)	0 (0%)	31

However, we classified an equal number (42%) of explanations as mathematical reasoning. In some cases, mathematical reasoning seemed central to answering the item correctly, as in Appendix item 2:

> I said (a), that (I) and (II) had the same kind of error. I worked them out again to see kind of what they were and when I did it I saw that they were carrying a one instead of a two in both cases. And for the third one, it was their addition, that they just had the wrong number in the ones place so they either added wrong or—. So I saw (I) and (II) were the same type of problem they had.

Here, teachers *must* perform a mathematical analysis in order to answer the problem correctly; in fact, teachers need not know whether this is a typical student error.

In another response by the same teacher, mathematical reasoning took place in conjunction with KCS:

> I said Jill [Appendix item 4] might not have known how to subtract eight from one so instead she did eight minus one is seven so she went down to the one and subtracted that way. And then . . . but she can subtract one from five so she got four. So I was just thinking that she didn't regroup it to give her that eleven to subtract the eight from and didn't check what she was doing because looking at it, it doesn't make sense. And I see kids do that all the time.

Here, we might argue that this teacher's ability to analyze the mathematical problem and familiarity with students' errors both facilitated her answer. For this same item, however, some teachers' answers appeared to rely more heavily on familiarity with the student error; others appeared to rely more heavily on the mathematical analysis. In each instance, such thinking usually led to a correct answer.

That teachers might use either KCS, mathematical reasoning, or both to get these items correct is not surprising. Logically, teachers must be able to examine and interpret the mathematics behind student errors prior to invoking knowledge of how students went astray. This multidimensionality was in fact reflected in our original specification of KCS as amalgamated knowledge of content and students, and it helps explain the factor analysis findings described above. Unfortunately, it also leads to problems of measurement, because items vary in the extent to which they draw on these two dimensions. We also find it particularly troubling that the cognitive interviews suggest that many of these problems can be answered in the absence of the KCS they were intended to assess.[2] This may help explain high teacher scores on this construct, as even if KCS is not present, mathematical reasoning and knowledge can compensate.

A third category of explanation for answers involved test-taking skills. Table 2 shows that test-taking strategies were prevalent in three of the five closed-ended items presented to respondents. In addition to providing evidence that these items tap unintended constructs, the use of test-taking skills here also explains the higher-on-average scores for the multiple-choice items. By crossing responses coded as "using test-taking skills" with the correctness of the response, we see that of 93 responses that were coded as test-taking, 76% resulted in a correct answer. For subject matter knowledge items, by contrast, test-taking strategies as often led to incorrect answers (49%) as correct answers (51%). This further amplifies the problems of measurement we face in this area; if the relatively common use of test-taking skills were not problem enough, the use of test-taking skills for KCS items leads respondents toward correct answers.

The cognitive interview data offer several other clues about problems of measurement in this domain. First, very few teachers selected outright "wrong" answers, or distractors, to problems such as the Bonny and decimal ordering items (Appendix items 1 and 3). In the MPDI pre-post sample, for instance, only 1% answered (a) and 4% answered (b) for the first; only 2% answered (c), and 1% answered (d) for the latter. The most common incorrect answer was the "all of the above" response,

[2] A related study (Hill, Dean, & Goffney, 2007) found that mathematicians and non-teachers tended to do quite well on this set of items via mathematical reasoning alone.

a pattern we saw repeated on many other items in our closed-ended survey. This suggests a hypothesis about the domain under study: that expert teachers are differentiated from nonexperts only by the specificity and detail of their interpretations of student problems. Teachers without such knowledge may accept many different explanations for student errors, be unable to discriminate common strategies, or be unable to identify developmental pathways—and thus accept any plausible answer. For this reason, measurement may be difficult in the context of a multiple-choice format, which depends heavily upon item distractors matching individuals' misconceptions and faulty reasoning processes. Whether this is the case or not is a topic for future, more qualitative, research.

A problem with the decimal ordering item (Appendix item 3) suggests a further complication, however: that the knowledge we intended an item to capture was unevenly measured because teachers had other knowledge of student errors and developmental sequences that "interfered." In the original coding, we specified that answers coded "KCS" had to match the KCS intended to be measured by the item. In relaxing this assumption, we see that teachers' "other" KCS often played a large part in their reasoning process. In the decimal ordering item, for instance, some teachers commented that students sometimes make errors when they see a whole number (7 in choice (b)) mixed with decimals, making this option attractive. Although we did not intend to measure this "knowledge of students," it seems a legitimate defense of (b) as a correct answer. Among the interviewed teachers, in fact, none answered (c) on the basis of the common "ignore decimal point" error, although roughly half commented to interviewers that they were familiar with this error. With the Bonny problem (Appendix item 1), a similar problem emerged:

> This is my kids. Bonny thinks that two and twenty are the same. I put that down. And then I got to Bonny doesn't know how large twenty-three is, and then Bonny doesn't understand the meaning of the places. And see, all of the above, for my kids, that's true. . . . I said all of the above.

This teacher's comments suggest either that she is a nonexpert who cannot distinguish between incorrect and correct explanations for Bonny's errors or that there is an element of truth in the distractors, based on her experience in the classroom.

These problems lead to a question: Can teachers' KCS be measured in multiple-choice format? Unfortunately, there is no "proof of concept" item, in the sense that factor analyses demonstrate it measures KCS; our interviews demonstrate teachers use KCS to answer it, the item statistics indicate it performs well in distinguishing among teachers, and it is reasonably difficult. However, our interview covered only six KCS items; our larger item pool contains several items that meet at least the psychometric criteria and a few that perform exceptionally well. Further cognitive interviews will help clarify whether these items do, as we suspect, better measure KCS.

An inspection of these successful items may provide hints as to future directions for measure development. By successful, we mean those that discriminate among teachers of different knowledge levels, and in particular those that discriminate

among highly knowledgeable and very highly knowledgeable teachers, where our scales were weakest. One type of highly discriminating and at least moderately difficult item asks teachers to inspect different student productions, and then make a determination about which knowledge was either a) most advanced, developmentally or b) reflected understanding of the mathematics, rather than procedural or factual recall. Another set of successful items asks teachers to determine which computational strategies students would be likely to use or not use in modeling problems or remembering basic facts. A third set of successful items asks teachers to supply the reasons for common student misconceptions or procedural errors. These all seem promising directions for future item development.

Results from the cognitive interviews suggest lingering problems with our conceptualization of this domain, including the need to answer questions surrounding whether it consists of recognition, recall and/or reasoning, and whether the research base or teachers' own non-research-based KCS should predominate; we discuss these at more length in the discussion and conclusion. Results from the cognitive interviews, however, do suggest explanations for the lower reliabilities and test information curves reported in the psychometric analysis: teachers' reliance on mathematical reasoning and test-taking strategies might have inflated scores. Results also suggest that KCS may be at least a portion of what has been measured by this set of items. With this in mind, we turn next to the pre-post MPDI analysis as a final check on validity.

Demonstrating Convergent/Discriminant Validity

Another approach to validity examines the associations between a set of measures and other constructs to which they are hypothesized to relate (convergent validity) and not relate (discriminant validity). Our pre-post analysis of the MPDI data provided the opportunity to take such an approach. We asked whether any growth in teachers' KCS scores related to teacher reports of covering KCS-related topics in their institute, rather than their reports of covering subject matter knowledge more generally.

Overall, teachers gained .21 standard deviations between pretest and posttest on the KCS measure. A t test showed that this gain is statistically significant (different from zero) at $p < .0001$. Substantively, this gain corresponds roughly to a 1–2 item increase in the raw number correct on our assessment. This average, however, masks variation in MPDIs; teachers in some institutes garnered 2-item gains, whereas others did not gain at all. An analysis of variance showed that these institute-level differences were significant ($F = 3.21$, $p < .0001$). Results from an unconditional mixed model (described below) suggest that 9% of the variation in gain scores lay between institutes, rather than between teachers within institutes.

Our cognitive interview results suggest that the overall growth exhibited on these measures might have resulted from improvements in teachers' actual KCS, or alternatively, in teachers' subject matter knowledge. Improvements in either would have boosted a teachers' score, assuming results from the cognitive inter-

views hold across the item set. To investigate this, we ran mixed models to determine the relationship between MPDI foci and teachers' gain; if teachers' gain scores were more related to their reports of learning KCS than pure subject matter knowledge, this adds to the evidence that the measures do capture some aspects of KCS. We also controlled for the pre-post combination of forms that teachers took, because results from the factor and item analyses suggested multidimensionality in our forms. If there were imbalances across forms in the extent to which KCS items drew on the KCS vs. general factors, these pre-post dummy variables would identify and control for them.

Table 3 shows correlations among independent variables, and Table 4 shows results from the mixed model analyses. In Model 1, we see that teachers' pretest score is the best predictor of their posttest score. Length of the institute is also marginally significant, with longer institutes producing stronger teacher gains. Teachers' reports of institute focus on KCS were significant at $p < .05$, suggesting that when teachers studied how students learn, they performed better on our posttest. Finally, the variable representing institutes that used student work as part of the curriculum was not significant and close to zero. It is difficult to say what this means, given that this is a 1-item variable, and thus prone to error of measurement.

Table 3
Correlations Between Predictors in Pre-Post Analysis

	Focus on KCS	Studied student work	Focus on subject matter knowledge	Focus on proof, analysis, representation, communication	Length of MPDI
Focus on KCS	1.00 (426)	.53** (422)	.53** (400)	.44** (426)	–.05 (426)
Studied student work		1.00 (423)	.10* (397)	.29** (423)	.06 (423)
Focus on subject matter knowledge			1.00 (401)	.45** (401)	.35** (401)
Focus on proof, analysis, representation, communication				1.00 (427)	.13* (427)
Length of MPDI					1.00 (429)

* $p < .01$. ** $p < .0001$.

Because of the small number of institutes and the correlation between the KCS and subject matter knowledge focus variables ($r = .34$), we could not jointly enter both into the same model. Doing so would have allowed us to identify whether one, the other, or both were responsible for teachers' growth on this construct.

Table 4
Pre-Post Analysis of Learning in MPDIs

	Model 1	Model 2
Intercept	–.35	–.54
	(.24)	(.26)
Pretest score	.58	.58
	(.04)	(.04)
Focus on KCS	.13	
	(.06)**	
Studied student work	–.05	
	(.04)	
Length of MPDI	.14*	.12
	(.06)	(.07)
Focus on subject matter knowledge		.06
		(.07)
Focus on proof, analysis, representation and communication		.10
		(.07)
Group B	–.24	–.24
	(.16)	(.16)
Group C	.30	.29
	(.19)	(.19)
Variance components		
Institutes	.04	.04
Residual	.42	.43
Akaike Information Criteria (AIC)	877.0	883.8

Note. Numbers in parentheses are standard errors.
* $p < .10$. ** $p < .05$.

However, Model 2 shows the effect of the subject matter focus variable and a related variable—focus on proof, analysis, representation, and communication—on teachers' outcome scores, controlling for pretest scores. This latter variable was shown to significantly predict teachers' growth in subject matter knowledge (Hill & Ball, 2004) and captures teachers' reports of the MPDI emphasizing these aspects of mathematical work. That neither it nor the subject matter focus predictor are significant here suggests that these scales have measured, in the context of the MPDI evaluation, growth in teachers' KCS.[3] Although there are indications of significant problems in the measurement of teacher knowledge in this area, the items and scales we developed are at least moderately sensitive to teachers' reported opportunities to learn KCS.

[3] More evidence for this contention is provided by the model fit statistic AIC, which is slightly lower for Model 1 than Model 2. A lower AIC indicates better model fit.

DISCUSSION AND CONCLUSION

Results from our analysis of the KCS items suggest several conclusions. First, we found that teachers do seem to hold "knowledge of content and students." The factor analyses of multiple forms and interviews with teachers suggest that familiarity with aspects of students' mathematics thinking, such as common student errors, is one element of knowledge for teaching. To those interested in identifying a knowledge base for teaching that might guide the preparation of teachers or the content of professional development, verifying that such knowledge is distinct from pure content or pedagogical knowledge is an important first step. Although it remains to be seen whether and how such knowledge, as we have measured it, is related to improving student learning in mathematics, our results bolster claims that teachers have skills, insights, and wisdom beyond that of other mathematically well-educated adults.

We also learned, however, that measuring such knowledge is not a straightforward enterprise. Although teachers' pre-post MPDI scores were responsive to reports of MPDI focus on students and content, results from other psychometric and validation analyses were much more equivocal. Scales, on the whole, were not nearly reliable enough to justify dissemination and use in research. Scales were poorly targeted, measuring less-than-knowledgeable teachers sufficiently but few others well. Both psychometric work and teacher interviews suggested that within the KCS domain, there was strong multidimensionality. Multidimensionality was, of course, part of our specification of this domain. However, the type of multidimensionality that emerged—with items relying in different amounts on mathematical reasoning, knowledge of students, and perhaps even on a special kind of reasoning about students' mathematical thinking—yields difficulties in the construction of parallel forms and teacher scores.

Overall, the effort described here resulted in a glass half full. More important, however, it also resulted in key lessons for future efforts to measure teacher knowledge in this and similarly structured domains. We consider three types of lessons: for conceptualization, for measurement, and for the criteria we proposed for undertaking both.

First, we learned that this domain remains underconceptualized and understudied. Although most scholars, teachers, and teacher educators would agree that teachers' knowledge of students' thinking in particular domains is likely to matter, what constitutes such "knowledge" has yet to be understood. Does KCS consist of facts that the research literature has uncovered about student learning? Of observations that experienced teachers make about student learning and errors? If the former, the research base is far from complete; we know far more about students' knowledge and thinking in areas such as fractions, early arithmetic, functions, probability, and geometry than we know in domains such as measurement, integers, and number theory. If the latter, the situation is similarly complicated. Research in educational psychology may miss important aspects of knowledge that teachers develop from their work with learners. Consequently, our items, by drawing on that research, may

have asked questions that do not resonate with teachers' experience and may have missed posing questions that tap the professional knowledge that teachers hold implicitly. Mapping this knowledge is likely to be a long and time-intensive process; however, it is necessary for the development of these and similar items.

Further, the very notion of "knowledge of content and students" as *knowledge* needs further development. Teachers "know" that students often make certain errors in particular areas, or that some topics are likely to be difficult, or that some representations often work well. But teachers also *reason* about students' mathematics: They see student work, hear student statements, and see students solving problems. Teachers must puzzle about what students are doing or thinking, using their own knowledge of the topic and their insights about students. Some teachers may do this with more precision and skill than others, but we lack theory that would help us examine the nature of teachers' mathematical-pedagogical reasoning about students. Some teachers may "know" things about students in particular domains, whereas others may be figuring out students' thinking as they teach. Thus we are probably measuring different skills and knowledge for different teachers, even within the same items.

In sum, our work suggests that the conceptualization of the domain is far from straightforward. In our case, we attempted to build the ship while sailing it—by writing items to help push our conceptualization and definition forward. Even though this proved theoretically productive, it no doubt contributed to our difficulties. We suspect that other research groups, however, are sailing similar boats. Our advice is to think carefully about the nature of the domain to be measured, including close analysis of the types of knowledge and reasoning that teachers do in their work.

The second set of lessons we draw involves measuring knowledge in this and similar domains. First, as originally formulated by Shulman and in our own theoretical work, this knowledge is explicitly multidimensional; as operationalized in discipline-specific, student-specific items, it cannot help but be. Unfortunately, psychometricians have little good news under these circumstances. Users must plan to collect large datasets and employ complex scaling techniques. If measurement endeavors like these are to succeed, as we think they must, there is an urgent need for more capacity on both fronts.

Next, these results have led us to think carefully about the multiple-choice format. Results from our cognitive interviews suggest that test-taking and mathematical reasoning helped respondents arrive at the correct answer. There is also a possibility, not examined in these retrospective cognitive interviews, that the items "taught" the content, essentially causing an "aha" moment for many teachers. In fact, we failed to write—and teachers failed to select—outright "wrong" answers to our items. During writing, we found that answer choices we intended to be wrong often seemed absurd, even to those with little knowledge of students. Other "wrong" answers had a grain of truth to them, making it difficult to penalize respondents who selected them.

This suggests that a possible direction for future item development in this and similar domains might be to invest in open-ended items like Appendix item 4. Unfortunately, these measures require that they be scored by hand before scaling, a considerable expense for large-scale studies. And we argue that there is need for

measures for such studies: currently, educational economists are using value-added models to determine what makes effective teachers. It is vitally important for classroom-based research to contribute both conceptions of good teaching *and measures of* good teaching to this work. Although we did not find an item among the six studied intensively here that met all of our proof-of-concept criteria for the multiple-choice format, we do have promising items in our larger pool. Investigating these items via interview and open-ended task studies is critical.

Finally, this study suggests the importance of using explicit criteria to guide measure conceptualization and development. We could have easily written items, administered them to teachers, and reported that we detected MPDI success in a pre-post analysis. Instead, we used the criteria to uncover significant problems in both the conceptualization and measurement of this domain. Rather than see our effort end here, however, we chose to use our results to demonstrate the benefits of using such an approach and to note aspects of the domain that we believe will prove useful to other researchers and future studies.

REFERENCES

AERA/APA/NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ashlock, R. B. (2002). *Error patterns in computation: Using error patterns to improve instruction*. Columbus, OH: Merrill Prentice Hall.

Ball, D. L., Hill, H. C., & Bass, H. (2005, Fall). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator, 29*, 14–22.

Begle, E. G. (1972). *Teacher knowledge and student achievement in Algebra* (SMSG Study Group Reports #9). Stanford, CA: Stanford University.

Begle, E. G. (1979). *Critical variables in mathematics education: Findings from a survey of the empirical literature*. Washington, DC: Mathematical Association of America, National Council of Teachers of Mathematics.

Behr, M. J., Harel, G., Post, T., & Lesh, R. (1992). Rational number, ratio, and proportion. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 296–233). New York: Macmillan.

Bryk, A. S., & Raudenbush, S. W. (1988). *Hierarchical linear models*. Newbury Park, CA: Sage.

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

Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education, 19*, 29–37.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.-P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*, 499–531.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school*. Portsmouth, NH: Heinemann.

Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., et al. (1991). Assessment of a problem-centered second-grade mathematics project. *Journal for Research in Mathematics Education, 22*, 3–29.

Franke, M. L., Carpenter, T. P., & Levi, L. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal, 38*, 653–689.

Franke, M. L., & Kazemi, E. (2001). Learning to teach mathematics: Focus on student thinking. *Theory into Practice, 40*, 102–109.

Fuson, K. (1992). Research on whole number addition and subtraction. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 243–295). New York: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

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

Hill, H. C., Dean, C., & Goffney, I. M. (2007). Assessing elemental and structural validity: Data from teachers, non-teachers, and mathematicians. *Measurement: Interdisciplinary Research and Perspectives, 5*, 81–92.

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

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

Kamii, C. (1985). *Young children reinvent arithmetic*. New York: Teachers College Press.

Kane, M. (2004). Certification testing as an illustration of argument-based approach validation. *Measurement: Interdisciplinary Research and Perspectives, 2*, 135–170.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences, 13*, 129–164.

Lamon, S. J. (1999). *Teaching fractions and ratios for understanding: Essential content knowledge and instruction strategies for teachers*. Mahwah, NJ: Lawrence Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*, 125–145.

Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal, 40*, 905–928.

National Board for Professional Teaching Standards. (2001). *NBPTS early childhood generalist standards: Second edition*. Arlington, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

QSR International (2002). QSR N6. [Computer software]. Doncaster, Australia: Author.

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual basis of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education, 20*, 8–27.

SAS Institute, Inc. (1999). SAS Version 8.12. [Computer software]. Cary, NC: Author.

Saxe, G. B., Gearhart, M., & Nasir, N. S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. *Journal of Mathematics Teacher Education, 4*, 55–79.

Schilling, S. G. (2005). *ORDFAC: A computer program for ordinal factor analysis* [Computer software]. Ann Arbor: University of Michigan.

Schilling, S. G. (2007). The role of psychometric modeling in test validation: An application of multidimensional item response theory. *Measurement: Interdisciplinary Research and Perspectives, 5*, 93-106.

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

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323–355.

Sudman, S., Bradburn, N. M., & Schwartz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tirosh, D. (2000). Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions. *Journal for Research in Mathematics Education*, 31, 5–25.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73, 89–122.

VanLehn, K. (1989). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). Bilog MG-3. [Computer software]. Chicago: Scientific Software International.

Authors

Heather C. Hill, Harvard Graduate School of Education, 6 Appian Way #445, Cambridge, MA 02138; heather_hill@harvard.edu

Deborah Loewenberg Ball, University of Michigan School of Education, 610 E. University, Ann Arbor, MI 48109; deborahball@umich.edu

Stephen G. Schilling, University of Michigan School of Education, 610 E. University, Ann Arbor, MI 48109; schillsg@umich.edu

APPENDIX: SAMPLE ITEMS

1. You are working individually with Bonny, and you ask her to count out 23 checkers, which she does successfully. You then ask her to show you how many checkers are represented by the 3 in 23, and she counts out 3 checkers. Then you ask her to show you how many checkers are represented by the 2 in 23, and she counts out 2 checkers. What problem is Bonny having here? (Mark ONE answer.)
   a. Bonny doesn't know how large 23 is.
   b. Bonny thinks that 2 and 20 are the same.
   c. Bonny doesn't understand the meaning of the places in the numeral 23.
   d. All of the above.

2. Mrs. Jackson is getting ready for the state assessment, and is planning mini-lessons for students focused on particular difficulties that they are having with adding columns of numbers. To target her instruction more effectively, she wants to work with groups of students who are making the same kind of error, so she looks at a recent quiz to see what they tend to do. She sees the following three student mistakes:

$\overset{1}{}$	$\overset{1}{}$	$\overset{1}{}$
38	45	32
49	37	14
+ 65	+ 29	+ 19
142	101	64
(I)	(II)	(III)

   Which have the same kind of error? (Mark ONE answer.)
   a. I and II
   b. I and III
   c. II and III
   d. I, II, and III

3. Mr. Fitzgerald has been helping his students learn how to compare decimals. He is trying to devise an assignment that shows him whether his students know how to correctly put a series of decimals in order. Which of the following sets of numbers will best suit that purpose? (Mark ONE answer.)
   a. .5 7 .01 11.4
   b. .60 2.53 3.14 .45
   c. .6 4.25 .565 2.5
   d. Any of these would work well for this purpose. They all require the students to read and interpret decimals.

4. Consider Jill's response to a subtraction problem. How might she have gotten an answer like this?

$$\begin{array}{r} 51 \\ -\ 18 \\ \hline 47 \end{array}$$