Journal for Research in Mathematics Education
2004, Vol. 35, No. 5, 330–351

Learning Mathematics for Teaching: Results from California's Mathematics Professional Development Institutes

Heather C. Hill and Deborah Loewenberg Ball
University of Michigan

Widespread agreement exists that U.S. teachers need improved mathematics knowledge for teaching. Over the past decade, policymakers have funded a range of professional development efforts designed to address this need. However, there has been little success in determining whether and when teachers develop mathematical knowledge from professional development, and if so, what features of professional development contribute to such teacher learning. This was due, in part, to a lack of measures of teachers' content knowledge for teaching mathematics. This article attempts to fill these gaps. In it we describe an effort to evaluate California's Mathematics Professional Development Institutes (MPDIs) using novel measures of knowledge for teaching mathematics. Our analyses showed that teachers participating in the MPDIs improved their performance on these measures during the extended summer workshop portion of their experience. This analysis also suggests that program length as measured in days in the summer workshop and workshop focus on mathematical analysis, reasoning, and communication predicted teachers' learning.

Key words: Elementary, K–8; Evaluation; Policy issues; Professional development; Teacher knowledge

Widespread agreement exists that mathematics education in the United States needs improvement. Over the past decade, a spate of resources have been developed to spur such improvement, including curriculum materials, frameworks, standards and assessments, and professional development. Efforts to make use of these resources have made clear that teachers' knowledge of mathematics is crucial for improving the quality of instruction (Ball, 1991; Ma, 1999; Sherin, 1996). The work that teachers are expected to do—use new curriculum materials, open their classrooms to wider mathematical participation by students, help students succeed on more challenging assessments—demands substantial mathematical skill. Teaching

Development of the measures used in this study was supported by NSF grant REC-9979873 and by a subcontract to CPRE on Department of Education (DOE), Office of Educational Research and Improvement (OERI) award #R308A960003. The development of the evaluation was supported by UCOP grant #8047PCE186, and the analysis of results supported by NSF grant REC-0207649. We would like to thank Patrick Callahan, Rena Dorph, Peter Goldschmidt, David Goldstein, Delena Harrison, Brian Rowan, Steve Schilling, Elizabeth Stage, Molly Vitorte and three anonymous reviewers for their assistance with this project. Errors are our responsibility.

mathematics requires an appreciation of mathematical reasoning, understanding the meaning of mathematical ideas and procedures, and knowing how ideas and procedures connect. Evidence that U.S. teachers' mathematical knowledge is unevenly developed (e.g., Ball 1990; Ma, 1999) has led to an increase in opportunities for teachers to learn mathematics for teaching.

One significant example was California's Mathematics Professional Development Institutes (MPDIs). The MPDIs were one component of the California Professional Development Institutes, a statewide program designed to provide subject matter teacher professional development in English language arts, English language development, and mathematics. Initiated in 2000, the mathematics program involved both mathematicians and mathematics educators in the design and implementation of content-focused, extended learning opportunities for teachers. The MPDI program was no small intervention: California educators undertook solving the problem of teachers' mathematical content knowledge on a grand scale, serving over 23,000 K–12 teachers in the first 3 years of the program.[1] This makes the MPDIs the largest content-focused professional development program in the United States, and it involved a significant expenditure of public funds.

One problem facing programs such as the MPDI, however, has been the lack of evidence about what constitutes effective professional development to improve teachers' content knowledge for teaching mathematics. This lack is related to the scarcity of reliable and valid instruments to assess program outcomes. Until recently, many instruments aimed at measuring teachers' mathematics knowledge—such as licensing exams—tapped computational ability rather than content knowledge as it is used in teaching (e.g., the ability to represent numbers or operations; the ability to respond to unusual student responses; see Ball, 1990; Hill, Schilling, & Ball, 2003; Ma 1999; Shulman, 1986). Without the latter type of measures, programs focused on improving teachers' mathematical knowledge for teaching could not evaluate their own success appropriately. Moreover, scholars, policymakers, and others knew little about whether and how teachers' mathematical knowledge was improved in particular programs—such as professional development or teacher preparation programs—or through other opportunities—such as using curriculum materials, participating in lesson study, or learning from practice.

This article chronicles our first efforts to use an instrument designed to measure teachers' content knowledge for teaching as an evaluation tool. We ask whether elementary school teachers can learn mathematics for teaching in a relatively traditional professional development setting—the summer workshop component of California's K–6 MPDIs—and if so, how much and what those teachers learn. We demonstrate how novel measures of teachers' content knowledge for teaching can be deployed to evaluate a large public program rigorously. And we set the stage

[1] This number includes those who participated in MPDI-run institutes funded by both the MPDI and the English Language Development PDI funding.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

for future analyses of the *conditions under which* teachers learn mathematical content by exploring the degree to which individual and institute-level characteristics might contribute to growth on our measures.

BACKGROUND

We begin with some background and a discussion of the methods underlying our analyses. Early efforts to study the relationship between teachers' mathematical knowledge and student achievement used characteristics of teachers and their educational background, such as courses taken, degrees attained, or certification status, as proxies for ability (Begle, 1979). Despite disappointing findings—that is, there was little relationship between these measures and student achievement—interest has persisted in establishing the connections between teacher knowledge and student achievement. By the mid 1980s, scholars began to see the problem differently, reframing the question of how teachers' content knowledge might contribute to student learning by focusing on forms of knowledge more closely related to teaching (Ball, 1991; Kennedy, 1997; Ma, 1999; Shulman, 1986; Wilson, Shulman, & Richert, 1987).

This new notion of *pedagogical content knowledge* represented a shift in conceptions of teacher knowledge. One implication was that the content knowledge needed for teaching is connected to the specific problems of teaching elementary school content such as fractions, word analysis, or writing. The work cited above also suggested that at least in mathematics, *how* teachers hold knowledge may matter more than *how much* knowledge they hold. In other words, teaching quality might not relate so much to performance on standard tests of mathematics achievement as it does to whether teachers' knowledge is procedural or conceptual, whether it is connected to big ideas or isolated into small bits, or whether it is compressed or conceptually unpacked (Ball, 1990; Ball & Bass, 2000; Ball, Lubienski, & Mewborn, 2001; Ma, 1999). This work further suggested that student learning might result not only from teachers' content knowledge but also from the interplay between teachers' knowledge of students, their learning, and strategies for improving that learning.

Working within this framework, researchers in mathematics education focused increasingly on teachers' knowledge of mathematics for teaching. Ball and Bass (2000), Lamon (1999), Leinhardt and Smith (1985), Ma (1999), Simon and Blume (1994), and Thompson and Thompson (1994) studied teacher knowledge in particular topic areas such as fractions, multiplication, division, and rate. These studies helped to illuminate what *knowing mathematics for teaching* requires. For example, in teaching students to reduce fractions, teachers need more than the procedure. In addition, they need to be able to explain why that procedure works and what it means (Leinhardt & Smith 1985). Teachers also need to appraise student methods for solving computational problems, and when students use novel methods, be able to determine whether such methods would be generalizable to other problems. We call this knowledge *specialized knowledge of content*, based on our belief that it is unique

to individuals engaged in teaching children mathematics. We contrast this specialized knowledge with what we call *common knowledge of content*—being able to compute 35 × 25 accurately, identifying what power of 10 is equal to 1, solving word problems satisfactorily, and so forth. This common knowledge is not unique to teaching; bankers, candy sellers, nurses, and other nonteachers are likely to hold such knowledge. We argue that together, specialized and common content knowledge compose the content knowledge used in teaching mathematics. We believe that teachers of mathematics need *both* types of content knowledge to teach this subject matter competently.

Researchers began to develop new measures of teachers' knowledge of mathematics that probed the pedagogical dimensions of the content. In mathematics, they developed open-ended and multiple-choice prompts that asked teachers to explain the reasons for particular mathematical procedures or rules, design mathematical tasks to highlight key concepts, and reason about relationships and number (Ball, 1988, 1990). These measures located teachers' mathematical knowledge in particular tasks of teaching. For example, of interest was not simply whether teachers know that 1 3/4 ÷ 1/2 = 3 1/2, but whether they could explain what the expression means, could construct a concrete representation or word problem that corresponds to it, and could manipulate the elements of that model to show a solution. Using such measures, scholars also began to investigate how teachers' mathematical knowledge, construed in these more pedagogically attuned ways, contributes to student achievement. Rowan, Chiang and Miller (1997), for instance, identified teachers' mathematical content knowledge as a predictor of student achievement in 10th grade. Scholars have also documented the mode and variation in teachers' command of knowledge for teaching elementary school mathematics (e.g., Ball, 1988; Ma, 1999; Simon, 1993; Simon & Blume, 1994; Tirosh & Graeber, 1990; for a review, see Ball, Lubienski & Mewborn, 2001).

To date, however, none of these measures has been employed to understand how such useful and usable knowledge of mathematics *develops* in teachers. Instead, scholars have focused efforts on detailing what teachers know and do not know, what might be known by teachers in order to teach, and how U.S. teachers compare with those abroad (e.g., Ball, 1990; Ball & Bass, 2000; Borko, Eisenhart, Brown, Underhill, Jones, & Agard, 1992; Ma, 1999; Thompson & Thompson, 1994). However, given the current volume of projects designed to support teacher knowledge, such as the NSF/DOE Math-Science Partnerships, and the development of statistically reliable measures of such knowledge in the field of elementary mathematics (Hill, Schilling, & Ball 2003; Rowan, Schilling, Ball, & Miller, 2001), we are able to begin tracking the development of teacher knowledge in this arena and identifying factors which contribute to such growth.

California's MPDI program is the first we have tried to study using these new instruments. The program was chosen because of its size as well as the ambitious scope of state policy and the stated emphasis on improving teachers' knowledge of mathematics. Sponsors of the program's institutes, typically teams of mathematicians and mathematics educators based at University of California campuses,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

were required to design programs aimed at improving teachers' content knowledge in mathematics and to participate in state-mandated assessments of those efforts. Teachers attended summer institutes of 1 to 3 weeks' duration (for a total of between 40 and 120 hours), taught by mathematicians and mathematics educators. The teachers participated in up to 80 hours of follow-up during the school year, and received stipends of approximately $1500 for their participation. Some institutes focused on elementary mathematics, others on middle and high school. They varied also in their mathematical focus, from number and operations to geometry to algebra. The number of contact hours in this professional development effort and the scope of the program in California made this a significant professional development effort and a substantial policy initiative.

The assessment on which we report below focuses only on the elementary number and operations institutes that took place in summer 2001. On our visits to the institutes, we found considerable variation in program content, depth, pedagogy, and quality. However, the programs also had common features. Typically, the mathematician associated with the project spent the morning covering an elementary mathematics topic, for example, long division, order of operations, prime/divisibility/multiples, or basic fractions. The mathematician might lecture, engage teachers in a series of activities designed to enhance knowledge, or lead a class discussion. In many but not all cases, this morning session was heavily keyed to mathematics as it is used in elementary classrooms, such as representing long division using manipulatives or stories. The afternoon sessions, typically led by mathematics educators, built on the morning's mathematics by forging links to practice and state standards or providing teachers with activities for classroom use. In contrast to other professional development programs observed by the first author, where mathematics was often pointed to but seldom explored (Hill, 2004; see also Wilson, 2002), these programs were full of mathematical content (for a similar observation, see also Madfes & Holtzman, 2001). The institutes contained, at least as initially observed, substantial opportunities for teachers to learn mathematics.

METHODS

Because one goal of this article is to explain how researchers might mobilize an ambitious, large-scale effort to evaluate teacher learning, we describe in detail the elements that comprised this effort. We discuss, in turn, the study context, the evaluation partnerships, key stages in the design and implementation of Year 1 of this study, our measures, and our data analysis techniques.

Study Context

In keeping with current calls for accountability throughout the educational system, the MPDI program required a pre- and postevaluation of teachers' learning as a result of participation in the mathematics institutes. However, at the program's inception, program designers and evaluators faced a challenge: how to measure

teachers' mathematical knowledge for teaching reliably and on a large scale. By teachers' *mathematical knowledge for teaching*, we meant not only their common content knowledge but also their specialized knowledge for teaching mathematics. By *large scale*, we meant several hundred teachers, more than could be reached through interviews. And by *reliably*, we meant using high quality scales that meet standards of educational measurement, including factor and item response analyses that show that the set of items written taps a coherent underlying construct and produces a scale with a reliability of .7 or more.

Prior to our study, however, few options existed for measuring teachers' content knowledge for teaching mathematics on a large scale. Many teacher licensing examinations (e.g., California Basic Educational Skills Test, National Teachers Examination) rely on problems similar to those found on the SAT or eighth-grade NAEP to measure teachers' mathematics content knowledge. These tests assess teachers' ability to solve problems, identify terms, calculate, and use formulas. They do not examine teachers' ability to unpack mathematical ideas, explain procedures, choose and use representations, or appraise unfamiliar mathematical claims and solutions—specialized knowledge of content—and thus, we argue, are incomplete in their scope. The only existing measures of teachers' specialized knowledge for mathematics teaching consisted of interviews and open-ended written responses or single multiple-choice items culled from early work by Ball, Post, and others contributing to the Teacher Education and Learning to Teach project (Kennedy, Ball, & McDiarmid, 1993). None of these measurement strategies was suitable for studying the MPDI program because they could not be implemented on a large scale, and little was known about their reliability. We at the Study of Instructional Improvement (SII), who are examining programs of school improvement in high-poverty urban elementary schools, were designing new measures of teachers' content knowledge for this study. We needed sites to pilot our new measures on a large scale so as to provide enough cases for statistical analyses.

The partnership that emerged between the MPDIs and the SII allowed us to pilot our items on a large scale and provided MPDI program officials a potentially useable set of measures, measures aligned more directly with their program content than other available instruments. However, conducting this study on a large scale also brought problems. Although program officials could use political leverage to convince MPDI providers to use the evaluation at their site, those sites were spread widely across a large state, making the development of the evaluation infrastructure and monitoring of particular sites difficult. Some providers, concerned about alarming teachers by what appeared to them to constitute a "test," declined to administer the evaluation. Other concerns included assuring teacher privacy, the length of time needed to complete the instrument[2] and, perhaps, reluctance to subject their

[2] Because this was a pilot, and we were unsure whether all items included would scale well, we nearly doubled the number of items typically used to create measures of this kind. This increased the length of the instrument considerably.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

site to scrutiny. Still other providers encountered technical difficulties as they attempted to use the database, teacher tracking system, and data contracting company arranged by MPDI officials. Because of the exigencies of the political system, which meant funding and program authorization arrived late in the process of program start-up, these systems had been constructed in the process of actually using them. Hence, problems (e.g., missing booklets) could not be anticipated and corrected ahead of time.

As a result, the data from the 2001 administration is not ideal. Out of an estimated 2,300 teachers served by the elementary MPDI program, we have data on only 398 teachers. To a large degree, this low number is because one institute serving roughly 1,500 teachers refused to participate fully, based on the directors' belief that teachers might be upset by having to complete the instrument. Although some individual sites within this particular institute did have teachers take both the pre- and postassessments, many subsites had teachers complete only the posttest.[3] Ten more sites, each serving far fewer teachers (roughly 25–50 teachers each), also failed to complete the assessments, some sites because the assessment forms were not ready at the program start date, and others for the reasons outlined above. Further, there was frequently a discrepancy between the official number of teachers paid to participate and the number of booklets that appeared at the firm hired to create a database of teachers' answers to our instrument. Because of consideration for individual teachers' rights to consent, the MPDI program could require individual institutes to administer the booklets, but could not force individual teachers to take the assessment. This means, then, that results associated with this study should be thought of as tentative, because sites or teachers had ample opportunity to select themselves into or out of this evaluation, perhaps on the basis of their perceptions of teacher learning within that site. At the same time, however, there is some possibility that these selections were more random than not, for we hypothesize the majority of refusals occurred for reasons other than a perception of teachers' failure to learn. Additionally, these data are less than ideal in a second respect. The best evaluation designs randomly assign program participants to treatment and control groups, allowing analysts to estimate true program effects. Good evaluation designs locate and evaluate the growth of comparison groups made up of individuals who are similar to participating teachers yet who elected not to engage with the treatment. We have neither kind of design here. This means that we cannot be absolutely sure that any effects found are a result of the MPDIs, rather than natural teacher learning over the course of time, or the effect of retaking a similar assessment. We deal with this issue further in our discussion of results.

[3] For data from this large site, researchers made judgments regarding whether to include subsites based on the number of teachers with complete pre- and postdata. If 15 or more teachers returned both booklets, the subsite was considered to have administered the assessment in full. If fewer than 15 teachers returned both booklets, the site was deleted from the analysis.

Measures of Mathematical Knowledge for Teaching

The items used to gauge content knowledge for teaching mathematics were written by the authors of this article, in collaboration with others working on the SII project. These items were grounded in common tasks of mathematics instruction, were designed to elicit both teachers' common and specialized knowledge of content, and were drawn both from the research literature (e.g., Ball, 1993a, 1993b; Carpenter, Hiebert, & Moser, 1981; Lamon, 1999; Lampert, 2001; Ma, 1999) and from writers' experiences teaching and observing elementary classrooms. In this article, we analyze only those items in the domain of elementary number concepts and operations, although forms given to teachers also included items meant to gauge content knowledge for teaching patterns, functions, and algebra, and knowledge of students and content.

Because items focus on the mathematical knowledge needed for teaching numbers and operations, the assessment is aligned with MPDI Elementary Number and Operations institute goals. However, the assessment was not designed to match the curriculum of any particular institute. Because of this orientation, we lost the ability to make definitive statements about particular institutes' efficacy because the size of any effects is correlated at least in part with the overlap between institute curriculum and the assessment (for arguments about overlap, see Barr & Dreeben, 1983; Berliner, 1979; Leinhardt & Seewaldt, 1981). However, this orientation does allow us to make statements about how well the broad program contributes to teachers' knowledge of mathematics for teaching.

In order to investigate the effectiveness of the MPDIs, we administered a preassessment to teachers at the beginning of the institute and a postassessment at the conclusion of intensive summer work. These assessments did not contain the same items; in order to reduce the possibility of teachers recalling or discussing their prior answers and of providers tailoring instruction to the assessment, we designed the pre- and postassessments with only a small number of common items (6 in the area of elementary numbers and operations content knowledge) and a larger number (approximately 16–20) of items specific to each form. This strategy required equating forms to control for the difficulty of items presented to teachers, a process we describe below.

As shown in Table 1, the numbers and operations scale on each form contained 13–14 stems and 23–26 items. Sample items appear in the Appendix. The differ-

Table 1
Information for the Content Knowledge (CK) Scales for Teaching Elementary Numbers and Operation

Scale	Number of stems	Number of items	Reliability
CK-Form A	13	26	.72
CK-Form B	13	24	.78
CK-Form C	14	23	.71

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

ence between stems and items arises because some stems, such the one in item 2 shown in the Appendix, contain multiple items beneath them. In particular, the work for Student A, Student B, and Student C constitutes multiple items within one stem ("Imagine you are working with your class . . ."). Other stems, such as the third item in the Appendix, have only one item. In addition to items measuring teachers' content knowledge for teaching, each form also contained items meant to gauge other constructs we saw as being important to this analysis: factors that motivated teachers to participate in MPDIs, teachers' perceptions of MPDI mathematical content coverage, and teachers' perceptions of how content was covered—in what depth, with an emphasis on particular mathematical practices, and so forth. In contrast to the content knowledge for teaching items, which could be scored as correct or incorrect, these measures of motivation and MPDI content were constructed in conventional Likert format, which asked whether teachers agreed or disagreed with statements such as "The Mathematics Professional Development Institute will help me learn the mathematics I need to know for teaching," an item included in the "motivation to attend" set.

An important part of data analysis was developing scales from both the Likert and content knowledge for teaching mathematics items contained on our forms, and we provide details on this process here. The main goal of scale-building is to identify a set of items that help represent an individual's position on some unobservable trait, and then to assess with what precision the items make that estimate. With motivation, for instance, we were interested in whether any of the set of six items written to represent teacher's various reasons for attending the MPDI captured a construct we had intended to be *desire to learn*. A factor analysis of all motivation items suggest that three (including the one above) represent teachers' desire to learn within MPDIs yet these three only had an average internal reliability of .58. This suggests that the three items are insufficient, by conventional standards, to accurately represent individuals' position on this construct. Another way of saying this is to note that since reliability is defined as the proportion of true score to true score and error combined, this particular measure is error-filled, and cannot be counted on to accurately discriminate between teachers with different levels of desire to learn. We completed this process with other Likert items of teachers' motivation and perception, and present the results of one key scale below.

We began data analysis on the content knowledge items by performing a factor analysis on all items from each of the forms,[4] looking for items that together would represent a single construct, or ability, in teachers' knowledge. This factor analysis was more complex than those from Likert items, since methods had to be used to factor analyze dichotomously scored items where stems had multiple items nested beneath (e.g., item 2 in the Appendix). Results from this specialized factor analysis

[4] Teachers cycled through the three forms as pre- and posttests. One group, for instance, took Form A as a pretest, Form B as a posttest, and recently completed Form C after the year of program follow-up. Another group took Form B as pretest, Form C as posttest, and Form A as the follow-up test, etc.

suggested a strong content knowledge dimension running across items written to represent both number concepts and operations (Hill, Schilling, & Ball, in press). We took these items and, for each form, fit item response theory (IRT) models to confirm and learn more about these number concepts and operations content knowledge scales (see Hambleton, Swaminathan, & Rogers, 1991 and McDonald, 1999 for additional information on IRT). Additional details on the results from the factor analyses and IRT modeling, including information on specific factor structures and the number of misfitting items, can be found in Hill, Schilling, and Ball (2003).

To enable pre- and postcomparative analysis, forms were linked using common item equating methods. The need for equating arises in situations where individuals take different pre- and posttests in which one form of the test may be more difficult (because of having a selection of more difficult items) than the other. To control for this possibility, analysts use the items common to both assessments to gauge the relative difficulty of the rest of the items on the test. We did this for our study, using pretest data to equate forms and arrive at an estimate of the forms' relative difficulty vis-à-vis one another, and then correcting for imbalances in that difficulty. However, because of modest sample sizes on each of the pretests, we were not able to equate forms using a two-parameter IRT model. Instead, we used a one-parameter Rasch model that, after deleting two badly misfitting items, produced the scales in number and operations content knowledge across the three forms listed in Table 1. The interpretation of the reliability estimate is similar to the Likert case described above, although the reliability itself is much better (.78 instead of .58). The intuitive explanation of reliability is the proportion of "signal" to "signal plus noise" in the data, this scale captures more "signal" than the motivation scale. Conventional scaling rules of thumb suggest that reliabilities of above .70 are required to detect moderate effects in large groups, and we appear to have met that criterion here.

One additional wrinkle pertains to the scoring of our content knowledge for teaching mathematics scales and interpretation of the results. We did not score the forms in terms of raw score or percentage correct because percentage correct is not a linear measure—the difference between 10% and 20% is a more substantial difference in true ability than the difference between 50% and 60%. Instead, Rasch models report performance in terms of a linear measure called *logits*. Logits are equal-sized units, and range in our data from –3.0 to 5.0, with higher scores indicating teachers with more mathematical knowledge for teaching.

We believe the final scales reported in Table 1 represent teachers' knowledge of mathematics as used in teaching. The scales include items that both assess teachers' common knowledge of content (e.g., knowing that the number 8 can be written as 008; see item 1 in the Appendix) and their specialized knowledge of content for teaching. These specialized tasks included representing numbers and operations using materials or stories, providing reasons and explanations for concepts and algorithms, and making mathematical appraisals of students' work. We attempted to balance items on each form across grade level, across whole numbers (45% of

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

items), fractions (42%), and decimals (16%), and across number concepts (54%) and operations (46%).

Finally, our work to assess the validity of our measure is ongoing. A best-case investigation of validity would include comparing teachers' score on our content knowledge for teaching mathematics scale with an assessment of their use of mathematics content in actual classroom teaching; this work is currently under way. Less convincing, although more often done in the field of test construction, are cognitive tracing interviews, in which individuals talk through their thinking and answers to particular items. If respondents answer a particular item correctly, but explain it incorrectly, then problems of validity are likely. This is also true if correct mathematical reasoning underlies incorrect answers. An analysis conducted with similar items to these suggested that for knowledge of content items, teachers' answers did in fact represent their underlying reasoning process (see Hill, 2002). Finally, we also studied content validity by comparing item topics to strands in *Principles and Standards of School Mathematics* (National Council of Teachers of Mathematics [NCTM], 2000) and California's *Mathematics Framework for California Public Schools, Kindergarten through Grade Twelve* (California State Board of Education, 2000). The K–5 NCTM's Standards, written at a relatively high level of generality, were well covered by the elementary number and operations items used here. The California standards, written at a greater level of specificity, suggested that we should include more items to represent K–2 subject matter (Seidel & Hill, 2003).

Sample and Data Analysis

At MPDI sites, we collected complete pretest and posttest data from 398 teachers in 15 institutes, or from an average of 26 teachers per institute. Overall, each institute served about 37 teachers. Because attrition is always a concern in these pre- and poststudies, we performed a *t* test to examine whether those who left their MPDI before completing a posttest were significantly weaker, mathematically speaking, than teachers who stayed. Pretest scores for the two groups were not significantly different according to this analysis ($p = .86$).

After IRT scaling and form equating was complete, data analysis proceeded first using simple descriptive statistics and exploratory ANOVAs, and then mixed models using Hierarchical Linear Models (HLM), a program written to accommodate multilevel data (i.e., teachers nested within institutes) (see Bryk & Raudenbush, 1988). Unfortunately, the absence of adequate longitudinal data on teachers' growth of content knowledge for teaching mathematics in these institutes prevents us from using the more powerful growth model analysis possible in HLM. Instead, we present here results from a covariate adjustment model, predicting teachers' posttest scores from pretest scores plus institute characteristics, and observe that Rowan, Correnti, and Miller (2002) argue that in their data, this model underrepresents the influence of inputs on outcomes in comparison to a growth model. Such might also be the case here, with the effect of institutes on teachers underestimated.

RESULTS

Across all institutes for which we have data, the average teacher score on the equated pretest was .47 logits,[5] and the average teacher score on the posttest was 1.06 logits, for a gain of .48 logits. The standard deviation of the pre- and post-IRT scales were 1.05 and 1.29 logits, respectively, making this gain somewhere between a third and half standard deviation in size and statistically significant at $p < .0001$. Substantively, this gain corresponds roughly to a 2–3 item increase in the raw number correct on our assessment. Considered in light of the fact the assessment was not designed to match the curriculum of any particular institute, which decreases the likelihood of finding positive teacher growth, this gain is a promising finding.

An analysis of variance (ANOVA) using institutes as the dependent variable found significant differences between institutes on a number of key measures. First, there was significant variation among institutes in the baseline mathematical knowledge of teachers as measured by percentage correct on the pretest ($p < .001$). Some institutes worked with, on average, teachers who scored a half standard deviation lower than the mean on the pretest; other institutes worked with teachers who scored a half standard deviation above the mean. Second, institutes were differentially effective in increasing teachers' content knowledge. Roughly one third of institutes had teachers who performed no better on the posttest than on the pretest; about half the institutes had teachers who gained between a third and two thirds of a standard deviation on the scale representing content knowledge for teaching; about one sixth of institutes had teachers who gained, on average, a standard deviation or more. In these highly effective institutes, teachers answered 5–6 more items correctly on the posttest than on the pretest. This difference in institute effectiveness is statistically significant at $p < .0001$. This finding suggests that overall positive program effects were not due to taking a posttest that is similar to the pretest, because if this were the case, we would expect to see similar growth in all institutes. The degree of variation in institute effectiveness also suggests that modeling these results using hierarchical linear modeling (HLM) might be fruitful.

An initial two-level HLM model using only teachers' posttest scores as an outcome and no independent measures confirms variation among institutes. In this model, labeled Model 1 in Table 2, 13.6% of the variation lay between institutes. To control for teachers' preexisting mathematical knowledge of teaching, we next entered teachers' pretest score as an individual-level predictor of outcomes. This within-institute effect of pretest on posttest score was highly significant ($b = 0.80$, $p < .0001$), and resulted in a reduction in the amount of within-institute variance by 47% as shown in Table 2, Model 2. Put differently, within each institute, teachers' pretest scores positively, significantly, and strongly predicted posttest scores, a common finding in evaluation and educational research studies.

[5] Winsteps, the program used for form equating, calibrates its scale on the average item difficulty, rather than average teacher ability. Thus, the average teacher score on the pretest was greater than zero.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 2
Hierarchical Linear Model (HLM) Predictions of Institute Outcomes

	Model 1	Model 2	Model 3	Model 4
For intercept, β_0				
Intercept (γ_{00})	1.0***	.99***	.98***	1.01***
	(.14)	(.14)	(.10)	(.09)
Length (γ_{01})			.33*	.36**
			(.14)	(.11)
Average institute pretest (γ_{02})			.45	
			(.43)	
Number of teachers in institute (γ_{03})			.006	
			(.009)	
Proof/analysis focus (γ_{04})				.76*
				(.32)
For pre- and postslope (β_1)				
Intercept (γ_{10})		.81***	.80***	.80***
			(.04)	(.04)
Variance components				
Institute mean (u_0)	.23***	.26***	.12***	.08***
Residual (r)	1.47	.78	.81	.81
Degrees of freedom				
Teacher level	397	396	393	394
Institute level	14	13	11	12

Note. The numbers within parentheses represent standard errors.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Because teachers' pretest scores were centered around the mean teacher within their institute, the intercept, or β_0 in this second model corresponds to the mean institute outcome, controlling for teachers' knowledge as measured on the pretest. Because the reliability of this intercept is relatively high (.87), we were able to model mean institute outcome based on teachers' reports of factors such as desire to learn and institute content, the length of summer institute (1, 2, or 3 weeks), the average pretest score of the group of teachers in the institute, and the number of teachers who completed both pre- and posttests within each institute. Because the small number of institutes included in this data set (i.e., 15 institutes) prevents jointly testing all these variables, we ran separate models for survey (teacher motivation and MPDI content) and nonsurvey variables (length, average institute pretest score, number of teachers in each institute). We also constrained the slope of pretest on posttest—or the amount by which teachers' pretest scores predicts their posttest scores—to be equal for all institutes (see γ_{10} in Table 2) on evidence that this relationship varied only marginally among institutes ($p < .03$). In the case of survey results, we deleted variables that did not meet conventional standards for significance from our model and controlled for the length of institutes, to prevent confounding between length and institute content, in the case that these characteristics were interconnected. Table 2, Models 3 and 4, show the results of this exploratory analysis.

Model 3 in Table 2 shows average institute pretest score ($\gamma_{02} = .45$) was a positive yet nonsignificant predictor of gains across institutes; the institutes with more able entering teachers, then, did not differ significantly from those with less able entering teachers in the amount added to the average teachers' score. Nor was the number of teachers that completed the posttest assessments ($\gamma_{03} = .006$) in an institute (which we used as a proxy for institute size) significant. Length of the institute ($\gamma_{01} = .33$ in Model 3) measured in weeks ($n = 1, 2$, or 3), was a significant predictor of institutes' effectiveness; teachers with more contact hours tended to learn more. This echoes findings from other studies of professional development (Cohen & Hill, 2001; Garet, Porter, Desimone, Birman, & Yoon, 2001) and reinforces the conventional wisdom that longer sessions can enable more learning (e.g., Corcoran 1995).

We constructed a variable from teachers' answers to the following questions about institute content based on a 4-point scale from Strongly Disagree to Strongly Agree:

a) The MPDI helped me understand how to analyze similarities and differences among mathematical representations, solutions, or methods.
b) The MPDI contained opportunities to learn about proof and justification in mathematics.
c) The MPDI contained opportunities to learn about mathematical communication and representation.
d) The MPDI has shown me classroom activities that can involve students in exploration and investigations.

When scaled, these items had an internal reliability of .77. When entered into the model as a variable that we called *opportunity to engage in mathematical analysis, reasoning, and communication* ($\gamma_{04} = .76$), this scale significantly and positively predicted institute outcomes as shown in Table 2, Model 4. Thus, MPDIs in which teachers reported more mathematical activities of this type were likely to influence teachers' knowledge for teaching mathematics positively. Other variables (e.g., teachers' desire to learn within institutes; mathematical content covered) were not significant, because of a true lack of effect, because of poor question design, and/or because of low reliabilities in scales created from these variables. We hope to report on the effects of an improved set of measures of institute content (e.g., qualities of the mathematics explored, focus on student thinking, extent of grounding in classroom work) within the next few years, as these measures are developed and results become available.

To ensure the veracity of results given such a small sample of institutes, we performed several conventional statistical checks of the models reported in Table 2. Posttest scores, at both the teacher and institute levels, were fairly normally distributed. Although teacher-level residuals were not strictly normal due to "holes" at certain ranges, their distribution did not suggest a violation of normality either. Based

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

on a residual analysis, we deleted one outlier institute, reran the models, and returned nearly exactly the same results. And the rate of convergence for models, itself a diagnostic (see Bryk & Raudenbush, 1988, p. 202), occurred relatively rapidly, at under 20 iterations for all models reported here.

DISCUSSION AND CONCLUSION

In this article we report on a first analysis of a novel instrument for measuring teachers' mathematical content knowledge for teaching. Because of its novelty, the analysis is limited. The fact that several of the measures were under development at the time of this study, combined with problems with data quality (both response rates of institutes and, to a more limited degree, individuals within institutes) and the fact our data capture only two time points per teacher mean that these results are suggestive rather than definitive. This is particularly true of null findings like the one on the motivation variable *desire to learn*. The low scale reliabilities for our survey-based independent measures mean that we cannot say definitively that any of our null results will hold in an improved analysis. Subsequent analyses should be undertaken that account for the effect on outcomes of both *teacher characteristics*, such as motivation, educational background, teaching methods, and *institute characteristics*. These institute characteristics should include not only those typically measured by studies of professional development—for example, length, collaboration, focus on classroom-relevant practices—but also those that describe how content is actually addressed, including the mathematical topics covered within an institute, the treatment of mathematical ideas, and the nature of the tasks in which teachers are engaged.

In future studies, researchers should also apply strengthened research designs to programs of this sort. Because we did not have a control or even a comparison group, we cannot be sure that teacher learning would not have occurred anyway or that it was not a result of taking a similar assessment twice. However, we think it doubtful that teachers would have learned mathematics for teaching simply by balancing their checkbooks and reading the newspaper; instead, professional development, curriculum material use, interactions with students, and mathematics course taking are the most likely precipitants of increases in teacher learning in our study. The teachers participating in the MPDIs were engaged in professional development, but were not participating in other professional development outside of the MPDIs during the summer institute portion of this program. Test-retest effects are more of a threat to our results. However, we attempted to mitigate this possibility by using two different forms of the assessment, then linking these forms through IRT equating methods. Results that show institutes were differentially effective in improving teachers' scores from pretest to posttest suggest this strategy was effective, since a positive effect from retaking a similar assessment would be of similar size across all institutes.

Despite these problems with our design and analyses, the results do suggest some important findings. Our results show that teachers *can* learn mathematics for

elementary school teaching in the context of a single professional development program. This alone is news: policymakers, mathematics educators, and others can successfully design programs that improve teachers' content knowledge for mathematics teaching, a goal named prominently in many of today's published reports, policy recommendations, and research programs (e.g., Achieve, 2002; Conference Board of the Mathematical Sciences, 2001; Kilpatrick, Swafford, & Findell, 2001; National Commission on Mathematics and Science Teaching, 2000; NCTM, 2000; RAND Mathematics Study Panel Report, 2002). We were less successful in modeling institute characteristics connected with improving teachers' outcomes. We discuss the implications of each below.

Policymakers and others might use the overall positive results from this study to help shape future programs intended to improve teachers' content knowledge in mathematics. In our visits to these institutes, we noticed potential explanations for the success of this program that were not captured in our attempts to model results. First, program staff focused mainly on mathematical content itself, rather than on how to implement or modify instructional activities without discussion of those activities' mathematical content (e.g., Hill, 2004; Wilson, Theule-Lubienksi, & Mattson, 1996). By foregrounding mathematical content, providers increased teachers' opportunities to learn that content. Second, teachers had opportunities to work together on elementary-level problems that addressed problems that arise when *teaching* mathematics content. Teachers in one institute, for instance, investigated the mechanics of the long division procedure by matching steps in the symbolic algorithm to steps taken when dividing with concrete manipulatives. Teachers in another institute analyzed word problems to better understand their mathematical structure and listened to and observed both verbal and visual explanations for why $10^0 = 1$.

A third observation centered on who was involved and what they brought to the quality of the institutes' work: The institutes were taught by mathematically knowledgeable individuals (e.g., university mathematicians and mathematics educators), and participants were well-paid volunteers. Finally, although we identified similarities among the institutes, we also noticed that institutes differed in tone and content. Moreover—and particularly promising for future investigations—different "curricula" for teachers sometimes nonetheless produced comparable results in teacher growth. In one observed institute, professional development providers spent the better part of a day proving addition and subtraction algorithms using scientific notation. Teachers appeared engaged and asked extensive mathematical questions. In another institute, teachers worked on classifying subtraction problems according to the situations they model (e.g., compare), watched a demonstration of the standard multidigit subtraction algorithm using base ten blocks, then analyzed student errors in multidigit subtraction. These two institutes gained roughly the same amount on our content knowledge for teaching scale. This suggests that there may be more than one viable approach to improving teachers' mathematical content knowledge.

We were less successful in modeling how teachers gained knowledge within these institutes. However, we did find a clear effect of length of institute on the size of the difference between pre- and postassessments, with longer institutes tending to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

have higher gains. This finding, though, does not mean that only longer institutes can increase teachers' knowledge. An inspection of institute length by gains finds several counterexamples: the institute that tied for the highest gains in our sample was of only 1 week's duration. Further, several other 1-week institutes each improved teachers' knowledge by the same amount as 3-week institutes. More remains to be understood about the interplay of institute structure, curriculum, and pedagogy, and provider preparation and qualifications.[6]

We also fell short in modeling institute gains based on the variables meant to represent teachers' motivations and perceptions of institute content. One of our scales, however, notably did positively affect posttest scores. This scale represents teachers' opportunities for learning as they engage in analyses of solutions and solution methods, prove answers and outcomes, and learn about mathematical communication and representation. This suggests that the more teachers engage with mathematics in ways that afford them opportunities to explore and link alternative representations, to provide and interpret explanations, and to delve into meanings and connections among ideas, the more flexible and developed their knowledge will be. It may even be that the extent to which teachers encounter mathematics as a domain in which reasoning and representation are central, and not simply as one comprising rules and routines, the more they learn.

Although preliminary, these findings reinforce growing signs that the focus and content of teachers' learning opportunities may have important effects on the development of their mathematical knowledge for teaching (Sowder, Phillip, Armstrong, & Schappelle 1998). Whereas many stress the importance of allocating sufficient time to and extending the length of professional development (e.g., Corcoran 1995), our finding that the opportunity to engage in mathematical analysis, reasoning, and communication can improve teacher knowledge suggests that curricular variables may play an equally important role in the quality and impact of professional development. In a recent study of professional development, Cohen and Hill (2001) found that professional development was most likely to affect teachers' practice when it was focused on particular content as represented by curriculum, as well as on mathematical ideas, and on students' thinking about that same mathematics, and on teaching that content. Our results point to the need to probe more carefully into the content of professional development and to identify curricular variables associated with teachers' learning.

Finally, this study demonstrates that evaluations using relatively more rigorous measures of outcomes can be organized and implemented on a large scale. Kennedy (1999) describes a chain of evaluation that moves from first-level approximations to indicators of student outcomes (e.g., classroom observations; standardized student assessments) to fourth-level indicators (e.g., testimony about the effects of policies or programs). The better the indicator, the closer it can be linked to the

[6] Table 2 shows that variation between institutes remains in both models ($p < .001$ for among-institute variation).

quality of classroom instruction and student achievement. Typically, professional development studies include only fourth-level indicators, asking teachers to report on the extent of perceived learning or change in practice as a result of the professional development encounter. However, Kennedy concludes that "teachers . . . may not be good judges of whether they have learned from a program" (p. 358). Some studies, such as Cohen and Hill (2001) and Garet et al., (2001) also include what Kennedy categorizes as third-level indicators, that is, nonsituated reports on classroom practices, but these measures are still at a distance from what teachers actually do, both because of problems of validity and reliability of reports (Mayer 1999) and because many of the terms used in such measures (e.g., exploration; problem solving) are highly open to interpretation. These third-level predictors also do not capture teachers' content knowledge; to the extent that improving content knowledge is a program goal, however, sole use of practice measures is problematic. By contrast, this study employs what Kennedy (1999) calls second-level predictors. We use data drawn from instruments that present teachers with particular situations and mathematical problems that arise in those situations, then infer facility in teaching from their ability to solve those mathematical problems in the context of practice.

Thus, although not a direct measure of teachers' classroom work with students, the dependent measure based on pretest and posttest results used here does have several advantages. One is that it does not rely on teacher perceptions of program effect; it also measures the knowledge needed to teach mathematics. It measures such knowledge more objectively than other indicators, which rely on teachers' perceptions of growth or practice as a result of professional development, and thus leaves less room for positive findings based on social desirability of responses. With the continued development of these and similar measures for other content areas, as well as for other teaching domains such as knowledge of students, and ability to make appropriate instructional decisions, we hope more evaluations of this kind will become possible.

REFERENCES

Achieve. (2002). *Foundations for success: Mathematics expectations for the middle grades* (Consultation draft). Washington, DC: National Academy Press.

Ball, D. L. (1988). *Knowledge and reasoning in mathematical pedagogy: Examining what prospective teachers bring to teacher education.* Unpublished doctoral dissertation, Michigan State University.

Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *Elementary School Journal, 90*(4), 449–466.

Ball, D. L. (1991). Teaching mathematics for understanding: What do teachers need to know about subject matter? In M. Kennedy (Ed.), *Teaching academic subjects to diverse learners* (pp. 63–83). New York: Teachers College Press.

Ball, D. L. (1993a). Halves, pieces, and twoths: Constructing and using representational contexts in teaching fractions. In T. P. Carpenter, E. Fennema, & T. A. Romberg, (Eds.). *Rational numbers: An integration of research* (pp. 157–196). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ball, D. L. (1993b). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *Elementary School Journal, 93*, 373–397.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on the teaching and learning of mathematics* (pp. 83–104). Westport, CT: Ablex.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

Ball, D. L., Lubienski, S., & Mewborn, D. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching, 4th ed.* (pp. 433–456). New York: Macmillan.

Barr, R., & Dreeben, R. (1983). *How schools work*. Chicago: University of Chicago Press.

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

Berliner, D. (1979). Tempus Educare. In P. Peterson and H. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 120–135). Berkeley, CA: McCutchan.

Borko H., Eisenhart, M., Brown, C. A., Underhill, R. G., Jones, D., & Agard, P.C. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education*, 23, 194–222.

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

California State Board of Education (2000). *Mathematics Framework for California Public Schools, Kindergarten through Grade Twelve*. Sacramento, CA: Author.

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, 29–37.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works*. New Haven, CT: Yale University Press.

Conference Board of the Mathematical Sciences (2001). *The mathematical education of teachers*. Washington, DC: Author.

Corcoran, T. B. (1995). *Helping teachers teach well: Transforming professional development*. New Brunswick, NJ: Consortium for Policy Research in Education.

Garet, M. S., Porter, A. C, Desimone, L, Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Lessons from a national sample of teachers. *American Educational Research Journal*, *38*, 915–945.

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

Hill, H. C. (2002). *Cognitive tracing study: Validity of items in representing teachers' thinking*. Ann Arbor, MI: University of Michigan Study of Instructional Improvement.

Hill, H. C. (2004). Professional development standards and practices in elementary school mathematics. *Elementary School Journal*, *104*, 215–31.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2003) *Developing measures of teachers' mathematics knowledge for teaching*. Ann Arbor, MI: University of Michigan Study of Instructional Improvement.

Kennedy, M. M. (1997). *Defining optimal knowledge for teaching science and mathematics* (Research Monograph 10). Madison, WI: National Institute for Science Education, University of Wisconsin.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis*, *21*, 345–63.

Kennedy, M. M., Ball, D. L., & McDiarmid, W. G. (1993). *A study package for examining and tracking changes in teachers' knowledge*. East Lansing, MI: National Center for Research on Teacher Education, Michigan State University.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

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

Lampert, M. (2001). *Teaching problems*. New Haven, CT: Yale University Press.

Leinhardt, G., & Seewaldt, A. M. (1981). Overlap: What's tested, what's taught. *Journal of Educational Measurement*, *18*, 85–95.

Leinhardt, G., & Smith, D. A. (1985). Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology*, *77*(3), 247–271.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Mahwah, NJ: Lawrence Erlbaum Associates.

Madfes, T., & Holtzman, D. J. (2001). *Independent evaluator California professional development institutes—mathematics: Year one report*. San Francisco, CA: WestEd.

Mayer, D. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*(1), 29–46.

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

National Commission on Mathematics and Science Teaching for the 21st Century. (2000). *Before it's too late: A report to the nation*. Washington, DC: U.S. Department of Education.

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

RAND Mathematics Study Panel Report. (2002). *Mathematical proficiency for all students: A strategic research and development program in mathematics education*. Washington, DC: RAND Corporation.

Rowan, B., Chiang, F., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, *70*(4), 256–284.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record*, *104*, 1525–1567.

Rowan, B., Schilling, S. G., Ball, D. L., & Miller, R. (2001). *Measuring teachers' pedagogical content knowledge in surveys: An exploratory study*. (Research Note S–2). Ann Arbor, MI: Consortium for Policy Research in Education, Study of Instructional Improvement, University of Michigan.

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.

Sherin, M. G. (1996). *The nature and dynamics of teachers' content knowledge*. Unpublished doctoral dissertation, University of California, Berkeley.

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

Simon, M.A. (1993). Prospective elementary teachers' knowledge of division. *Journal of Research in Mathematics Education*, 24, 233–254.

Simon, M. A., & Blume, G. W. (1994). Building and understanding multiplicative relationships: A study of prospective elementary teachers. *Journal for Research in Mathematics Education*, *25*, 472–494.

Sowder, J. T., Phillip, R. A., Armstrong, B. E., & Schappelle, B. P. (1998). *Middle-grade teachers' mathematical knowledge and its relationship to instruction*. Albany, NY: SUNY Press.

Thompson, P., & Thompson, A. (1994) Talking about rates conceptually, Part I: A teachers' struggle. *Journal for Research in Mathematics Education*, *25*, 279–303.

Tirosh, D., & Graeber, A. (1990). Evoking cognitive conflict to explore preservice teachers' thinking about division. *Journal for Research in Mathematics Education*, *21*, 98–108.

Wilson, S. M. (2002). *California dreaming. Reforming mathematics education*. New Haven, CT: Yale University Press.

Wilson, S. M., Shulman, L. S., & Richert, A. (1987). 150 different ways of knowing: Representations of knowledge in teaching. In J. Calderhead (Ed.), *Exploring teacher thinking* (pp. 104–124). Sussex, UK: Holt, Rinehart & Winston.

Wilson, S. M., Theule-Lubienksi, S., & Mattson, S. (1996, April). Where's the mathematics? The competing commitments of professional development. Paper presented at the annual meeting of the American Educational Research Association, New York.

Authors

Heather C. Hill, Learning Mathematics for Teaching Project, 1600A School of Education Building, 610 East University Avenue, University of Michigan, Ann Arbor, MI 48109-7632; hhill@umich.edu

Deborah Loewenberg Ball, 4119 School of Education Building, 610 East University Avenue, University of Michigan, Ann Arbor, MI 48109-7632; dball@umich.edu

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission

APPENDIX

Sample items

1. Ms. Dominguez was working with a new textbook and she noticed that it gave more attention to the number 0 than her old book. She came across a page that asked students to determine if a few statements about 0 were true or false. Intrigued, she showed them to her sister who is also a teacher, and asked her what she thought.

   Which statement(s) should the sisters select as being true? (Mark YES, NO, or I'M NOT SURE for each item below.)

	Yes	No	I'm not sure
a) 0 is an even number.	1	2	3
b) 0 is not really a number. It is a placeholder in writing big numbers.	1	2	3
c) The number 8 can be written as 008.	1	2	3

2. Imagine that you are working with your class on multiplying large numbers. Among your students' papers, you notice that some have displayed their work in the following ways:

Student A	Student B	Student C
35	35	35
× 25	× 25	× 25
125	175	25
+ 75	+ 700	150
875	875	100
		+ 600
		875

   Which of these students would you judge to be using a method that could be used to multiply any two whole numbers?

	Method would work for all whole numbers	Method would NOT work for all whole numbers	I'm not sure
a) Method A	1	2	3
b) Method B	1	2	3
c) Method C	1	2	3

3. Ms. Harris was working with her class on divisibility rules. She told her class that a number is divisible by 4 if and only if the last two digits of the number are divisible by 4. One of her students asked her why the rule for 4 worked. She asked the other students if they could come up with a reason, and several possible reasons were proposed. Which of the following statements comes closest to explaining the reason for the divisibility rule for 4? (Mark ONE answer.)

   a) Four is an even number, and odd numbers are not divisible by even numbers.
   b) The number 100 is divisible by 4 (and also 1000, 10,000, etc.).
   c) Every other even number is divisible by 4, for example, 24 and 28 but not 26.
   d) It only works when the sum of the last two digits is an even number.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission