Journal for Research in Mathematics Education
2010, Vol. 41, No. 5, 479–512

Measuring the Effects of Professional Development on Teacher Knowledge: The Case of Developing Mathematical Ideas

Courtney A. Bell
Educational Testing Service

Suzanne Wilson
Michigan State University

Traci Higgins
TERC

D. Betsy McCoach
University of Connecticut

This study examines the impact of a nationally disseminated professional development program, Developing Mathematical Ideas (DMI), on teachers' specialized knowledge for teaching mathematics and illustrates how such research could be conducted. Participants completing 2 DMI modules were compared with similar colleagues who had not taken DMI. Teacher knowledge was measured with multiple-choice items developed by the Learning Mathematics for Teaching project and open-ended items based on problems initially developed by DMI experts. After controlling for pretest scores, a hierarchical linear model identified statistically significant differences: The DMI group outperformed the comparison group on both assessments. Gains in teachers' scores on the more closely aligned measure were related to the degree of facilitator experience with DMI. This study adds to our understanding of the ways in which professional development program features, facilitators, and issues of scale interact in the development of teachers' mathematical knowledge for teaching. Study limitations and challenges are discussed.

Key words: xxx xxxx xxx xxx, xxxxx xxx xxxxx, xxxx xxxx, xxxxx, xxxxx, xxxxx, xxxxxxx, xxxxxx

---

This article reflects a team effort. Authors contributed equally to the research and/or writing. This work was supported by the National Science Foundation under Grant No. ESI-0242609. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation or the staff of Developing Mathematical Ideas. The authors wish to acknowledge Young Oh for her contributions to the intellectual development of the open-ended assessment and scoring rubrics. We would also like to thank members of the DMI Advisory Board, especially Megan Franke for her contributions to the development of the open-ended assessment. Finally, we thank the staff of the Study of Instructional Improvement for their wise counsel concerning the assessment of teacher knowledge.

There seems to be little disagreement among policymakers, researchers, educators, administrators, and reformers that teachers are the critical component in improving U.S. education. Further, there is apparently universal agreement that high-quality, ongoing professional development for teachers is equally necessary (American Federation of Teachers, 2002; National Academy of Education [NAE], 2009). In addition, there appears to be consensus on the features of high-quality professional development, including,

- It focuses on deepening subject matter knowledge specifically for teaching, including understanding how students learn and the specific difficulties they may encounter in mastering key concepts.
- It involves enough time for significant learning (for example, a course or program of 40 or more hours distributed over 12 or more months).
- It is coherently related to what teachers are being asked to do and builds on what teachers already know and are able to do.
- Educators are actively engaged, rather than just listening to a lecture or watching a demonstration.
- Teams of teachers from the same school participate and learn together, enabling them to support each other in using what they have learned. (NAE, 2009, p. 6)

Although there is general agreement that these features are best practices, the empirical basis for making these claims is modest, and therefore additional research on the effectiveness of professional development programs is warranted (Desimone, Porter, Birman, Garet, & Yoon, 2002; Desimone, Porter, Garet, Yoon, & Birman, 2002; Garet, Porter, Desimone, Birman, & Yoon, 2001; Hill & Ball, 2004; NAE, 2009; Wilson & Berne, 1999; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). Effects are of two kinds: changes in teacher knowledge/practice and increases in student learning. Research linking professional development to changes in teacher knowledge or practice is suggestive. For example, Desimone, Porter, Garet, et al. (2002) found that "professional development focused on specific instructional practices increases teachers' use of those practices" and that the use of "active learning opportunities, increase the effect of the professional development on teachers' instruction" (p. 81). However, the research relied on teachers' self-reports of change and not on direct observations of practice.

Research that attempts to link professional development to K–12 student learning has produced mixed results. For example, Yoon, et al. (2007) examined nine studies that looked at the effect of teacher professional development on student achievement and found that the students of teachers who had, on average, 49 hours of professional development had greater achievement. But in an experimental study designed to test some of the aforementioned best practices, Garet, et al. (2008) found increases in both teacher knowledge and desired classroom practice in the short run. However, changes in teacher knowledge and practice did not translate into improved student achievement. Furthermore, the teacher effects were not sustained over time.

The empirical murkiness regarding the links among professional development, teacher knowledge, teaching practice, and pupil learning suggests that there is much we still do not understand about the complex interactions that make professional learning possible. Borko (2004) suggests that one way for the field to move forward is to conduct research that focuses on the interactions among facilitators, programs, and teachers in various sites. With such a focus, research would systematically and iteratively examine the capacity of professional development programs to produce changes in teacher knowledge, practice, and/or student learning in a single program in a single site (Phase 1), in a single program across multiple sites (Phase 2), and in multiple programs in multiple sites (Phase 3). The research reported here considers two issues critical to that research agenda: whether specific programs can be implemented at scale to produce desired outcomes, and which aspects of programs must be kept intact as they are scaled.

To our knowledge, the study reported here is the first Phase 2 professional development research study to examine changes in teachers' mathematical knowledge for teaching (MKT) in a single professional development program—Developing Mathematical Ideas (DMI)—across multiple sites with multiple facilitators. We consider two research questions: What MKT did participants acquire by participating in two commonly used modules of DMI? and How did sites structure DMI implementation?[1] We included the second question to explore potential explanations for cross-site variation. By considering these two questions, the study contributes to our understanding of professional development by providing one example of a program that was implemented at scale while producing evidence of change in teachers' MKT. The study also generates hypotheses for future research on scaling up and the methodological complexities of doing Phase 2 research.

DEVELOPING MATHEMATICAL IDEAS

Developing Mathematical Ideas is a widely used professional development curriculum designed for K–8 teachers of mathematics (e.g., S. Cohen, 2004; Schifter, Bastable, Russell, Cohen, et al., 1999; Schifter, Bastable, Russell, Yaffee, et al., 1999). DMI is widely used and is designed to align with many of the ideals expressed in the professional development literature. It takes seriously the importance of the development of specialized knowledge used in teaching mathematics, has anticipated the challenges of facilitation, and has provided materials with the potential to support the intensive, sustained, collaborative inquiry that is seen as crucial to effective professional development. Each of DMI's seven modules focuses on a different set of core mathematical ideas spanning grades K–8. Each module includes a facilitator's guide, a casebook for teachers, and video segments of elementary and middle school mathematics classrooms. Any given module is structured to fit within eight 3-hour sessions. The sessions are usually offered over a semester or

[1]Because the project was a modest one, and sites are spread across the country, it was impossible to also gather comparable data on student learning. It was also impossible to randomly assign teachers to the DMI and non-DMI groups.

during more intensive week-long summer sessions. The materials have also been adapted for use in preservice programs. Over 75,000 casebooks have been sold.

In DMI seminars, teachers work with a trained facilitator to learn about specific mathematics, learn about children's ideas about that mathematics, and analyze how to approach these ideas in the classroom. Each seminar focuses on a specific mathematical strand. During a typical session, teachers solve mathematical problems, discuss written and/or video-based classroom episodes, and examine students' work. All these activities revolve around a core mathematical idea. As teachers do mathematics together and discuss children's mathematical ideas, they learn to use a variety of representational devices and tools, communicate their ideas to others, and strengthen their own mathematical reasoning. The facilitator guides the process using a learner-centered, inquiry-oriented pedagogy for the teacher-participants. The written and video cases and samples of student work are designed to connect the mathematics that teacher-participants do in DMI seminars to the mathematics of elementary classrooms. To push the connection to teaching practice further, each session includes a homework assignment that engages teachers in writing about new ideas as they implement them in their own classrooms.

What teachers learn in DMI depends, in part, on the facilitator. Facilitators are supported in three different ways. First, they have access to module-specific facilitator guides that describe each session in detail, provide information on how to prepare for sessions, and describe plausible ideas and questions that may emerge during discussions. Second, the developers offer an intensive 2-week summer DMI Leadership Institute. Although such training is not required, many facilitators attend. During training, facilitators work through, with the help of experienced trainers, the modules they will teach. The emphasis is on both experiencing the DMI materials as a learner and analyzing that learning. A final resource for facilitators involves on-line support documents, including annotations and commentary about particular sessions from other facilitators.

BACKGROUND

As previously noted, there is a critical need for research that investigates whether and how professional development programs can be scaled up to support desired teacher learning. DMI, as a national professional development program that is stable and well articulated, is a reasonable site for doing such research. In particular, we focused on DMI's effects on teacher knowledge and, secondarily, on how variations in implementation might explain those effects. This second question helps inform the crucial issues of implementation and scale endemic to Phase 2 research that Borko (2004) identifies.

Measuring Teacher Knowledge

The technology available to researchers interested in measuring teacher learning is limited. Traditionally, professional development programs have asked teachers

to participate in exit surveys about a program's relative merits. Some professional development programs have attempted to use locally developed pre- and posttests of teacher outcomes. In general, professional developers have no warehouse of reliable, robust measures to use to assess changes in what teachers understand or can do.

The picture is equally bleak in teacher education research. Researchers have used a variety of different proxies for teacher knowledge and/or practice, many of which are relatively gross indicators: college majors, grade point averages, and retention (Cochran-Smith & Zeichner, 2005; Goldhaber & Brewer, 2000; Guyton & Farokhi, 1987; Monk, 1994; Wilson, Floden, & Ferrini-Mundy, 2001). Teacher tests are equally problematic (Wilson & Youngs, 2005). In the larger domain of research on teacher learning, instruments used to measure changes in teacher knowledge range dramatically in focus, quality, purpose, and utility (e.g., Kennedy, 1999; Porter, Youngs, & Odden, 2001; Seidel & Shavelson, 2007). Although variation in detail and perspective is desirable, there are few well-validated instruments. Thus, for professional development programs interested in collecting professionally responsible, publicly credible evidence of what teachers learn, there are few trustworthy methods or measures.

Because this research focused on DMI and because DMI focuses on developing teachers' knowledge of mathematics, children's ideas about mathematics, and instructional approaches that engage and support student reasoning, we were particularly interested in locating validated instruments for measuring elementary teachers' knowledge in those domains. Building on earlier work by Ball and other researchers in the National Center for Research on Teacher Education who explored ways to measure teachers' mathematical knowledge with a variety of item formats (i.e., Kennedy, Ball, & McDiarmid, 1993), researchers working on the Learning Mathematics for Teaching (LMT) project (e.g., Ball & Rowan, 2004; Hill, Ball, Blunk, Goffney, & Rowan, 2007; Hill, Rowan, & Ball, 2005; Hill, Schilling, & Ball, 2004) have developed survey items to measure teachers' mathematical knowledge for teaching (which we address subsequently) and have validated those items through a series of studies (Hill & Ball, 2004; Hill, Schilling, & Ball, 2004). We used selected LMT items to assess teachers' learning. Using these items provided us one way to assess the external validity of changes in teachers' knowledge that we documented with other local measures that we developed. We discuss these other measures in the next section.

LMT items are based on a conception of teacher knowledge that has also been explicated by LMT researchers (e.g., Ball, Thames, & Phelps, 2008; Hill et al., 2008) and further developed by other mathematics education researchers: mathematical knowledge for teaching. Earlier work on teacher knowledge included Shulman's proposal that teachers had both content knowledge and pedagogical content knowledge, as well as other forms of knowledge (e.g., Ball, 1989, 1990; Shulman, 1986, 1987; Wilson, Shulman, & Richert, 1987), as well as Ma's (1999) comparative study of Chinese and U.S. teachers, which captured the imagination of many mathematicians. In generating their "practice-based" theory of teacher knowledge, Ball and her colleagues proposed a new conceptualization that included

both content knowledge and pedagogical content knowledge: mathematical knowledge for teaching (MKT). This conception includes common content knowledge (CCK), specialized content knowledge (SCK), and "horizon" content knowledge on the content side and knowledge of content and students (KCS), knowledge of content and teaching (KCT), and knowledge of content and curriculum (KCC) on the pedagogical content knowledge side (Ball et al., 2008). We define these forms of knowledge later in the article when discussing the specific instruments used.

The LMT items are developed and validated using this map of teacher knowledge. For this study, items that aligned most closely with DMI included those targeting SCK, KCS, and KCT, as the majority of the seminars are focused on having teachers solve mathematical problems that are closely aligned with the elementary curriculum, engage in tasks that showcase children's thinking about mathematics, and implement instructional strategies in their classrooms. In other words, DMI focuses on three aspects of MKT, one that is in the subdomain of content knowledge and two that are in the subdomain of pedagogical content knowledge.

Scaling Up Professional Development

In addition to needing more research on the outcomes of professional development, the field also needs more knowledge about how to structure professional development programs so people other than the developers can successfully implement them at scale. Central here is the role of facilitators or professional development leaders. There exists little research that analyzes the characteristics or preparation of effective facilitators.

However, case-study evidence suggests facilitators can act as important resources for professional development (e.g., Stein, Silver, & Smith, 1998; Stein, Smith, & Silver, 1999). For example, anecdotal and case-based evidence suggest that effective facilitators need to articulate professional development program goals clearly to participants, create a trusting community of learners, know how to productively guide teachers toward deeper understandings, and adapt the program to local needs, while also not compromising program effectiveness. However, we have yet to establish a solid research base to support these suggested best practices (Borko, 2004; NAE, 2009). Although scaling up was not the focus of this study, our secondary research question concerned capturing variation in implementation across the sites should that variation account for the DMI effects we observed.

METHOD

The research reported here is based upon an NSF-funded evaluation of DMI that was designed to provide summative analyses on DMI's effects. The study involved 10 DMI sites, 9 of which included comparison groups. The research was designed to compare the MKT of a group of teachers who participated in DMI with that of a comparable group of teachers who did not. The design was nested, and although

we did not attempt to build a complete model of teacher learning using hierarchical linear modeling (HLM), we did use a two-level model that specifies teacher learning as a function of DMI and facilitator characteristics. Given our limited sample size, the model includes only one variable at the site level (level 2).

Participating Sites

Site selection involved several steps. First, we asked DMI staff to nominate sites where DMI had been implemented for some time. We selected stable sites because such sites would allow us to study the effects of DMI rather than the effects of initial DMI implementation. We explained the details of the research design and invited sites to participate. Each site had a local study facilitator who administered instruments. We eventually established relationships with 10 sites: Wareham, MA; Ft. Smith, AK; Bellevue, WA; Seattle, WA; Houston, TX; South Hadley, MA (2); Hudson, MA; Loveland, CO; and Bryant, AK. In the case of South Hadley, MA, there were two separate cohorts of teachers taught by two different sets of facilitators. Within the 10 sites there were two types: sites at which all were teachers from the same district and sites at which teachers came from different districts to a regional professional development center. Both South Hadley sites, Hudson, and Bryant were regional professional development centers.

Participants

All teachers were voluntary participants in DMI or peers of the DMI participants who volunteered to serve in the non-DMI comparison group. An experimental design utilizing randomization was not an option. The non-DMI group of teachers taught at the same site, but due to resources, was not matched by grade level, amount of teaching experience, or the like. By recruiting both groups from the same district, we attempted to hold district-level factors—such as concurrent reform initiatives, school demographics, and student curriculum used—constant. We administered a survey to all teachers in both groups to obtain information on background characteristics that might be relevant to teachers' professional knowledge or motivation to learn. The survey inquired about the hours and nature of mathematics-specific professional development in which teachers had engaged during the previous 3 years, courses taught, grade levels taught, highest educational level achieved, number of years as a full-time teacher, student population taught, gender, race/ethnicity, and so on. The DMI and non-DMI groups are described in more detail in the following paragraphs.

There were 308 teachers who participated. The majority were female (DMI 90%; non-DMI 87%) and Caucasian (DMI 81%; non-DMI 85%). The average number of years of teaching experience for DMI participants was 12.60 (SD = 8.75), and the average for non-DMI participants was 12.92 (SD = 10.28). The majority of participants were elementary school teachers, although a few participants were administrators/specialists or special education, middle school, or high school teachers.

As Table 1 details, DMI participants had a greater proportion of administrators and secondary teachers than non-DMI participants. Participants taught across a number of grade levels, with the majority teaching in grades 2–3 (21% DMI and 28% non-DMI) or in grades 4–6 (49% DMI and 46% non-DMI). Prior to the study, both DMI and non-DMI participants reported similar numbers of mathematics-specific professional development hours, with the exception of a small cohort of DMI teachers who had more than 81 hours of mathematics professional development in the past 3 years. We addressed these differences by using a professional development variable

Table 1
Participants' Educational Roles, Previous Mathematics Professional Development Hours, and Estimates of Students' Free and Reduced Lunch Eligibility (Percent)

	DMI	Non-DMI
Role		
Special education teacher, paraprofessional, bilingual or ELL teacher	10	9
Elementary teacher	62	77
Administrator, specialist or nonclassroom role	13	1
Middle or high school teacher	11	8
Hours of mathematics PD in past 3 years		
0	8	16
1–16	33	38
17–32	13	16
33–56	11	17
57–80	7	4
81+	23	4
Students' free and reduced lunch eligibility		
0%	1	2
10%	7	8
20%	4	14
50%	15	10
70%	9	5
90%	9	11
100%	6	6

Note. The response rate for educational roles was 96% and 95% for DMI and non-DMI groups, respectively. The response rate for professional development hours was 95% for both groups. The response rate for estimated free and reduced lunch eligibility was 49% and 44% for DMI and non-DMI groups, respectively.

as a covariate in our analyses. We also used pretest scores to control for potential initial differences between the average scores of the two groups on MKT.

Data on the schools in which participants taught was not consistently reported, but it appears that both DMI and non-DMI teachers taught in schools with varying percentages of students eligible for free or reduced lunch. In both groups, a majority of teachers worked with student populations in which less than 50% received free or reduced lunch.

Site Sample Description

The original sample consisted of 308 teachers from 10 sites; 9 of the 10 sites ($n = 257$) contained both treatment and comparison teachers. The 10th site ($n = 51$) was a training-of-trainers site, which included only treatment teachers. We began data gathering at 11 sites. After we received pretest data from one of these sites, the site dropped out of the study. Attrition took three forms: teachers who dropped out of the professional development, teachers who dropped out of the study, and teachers who did not follow directions concerning using an ID number that they created for themselves at both pre- and posttest. Each of these events resulted in unmatched data that we classify here as attrition. For 6 of the remaining 10 sites, attrition rates were as expected, ranging between 2% and 17%, and fairly evenly distributed between the DMI and comparison groups. An additional site had attrition rates of about 33%, also equally distributed between groups. One site with an overall attrition rate of 26% lost most participants from the comparison group (100% of the DMI participants at this site completed the posttest compared to 55% from the comparison group). The remaining two sites had high attrition rates. At one of these sites, the attrition rate was 46%, but this was equally distributed across both groups. The final site also had problematic attrition rates; the majority of participants who were lost from this site were from the DMI group. At the sites that had higher attrition (26–46%), a large portion of attrition resulted from inability to match pretest and posttest data.

Thus, there was great variability in the final sample sizes across the 10 sites. Whereas the smallest site consisted of only 14 teachers, the largest site involved 66 teachers. Table 2 shows the sample sizes and the standardized means and standard deviations (z-scores) for the treatment and comparison groups by site for the preassessments and postassessments. Because Site 1 was missing teacher-level data on professional development and other teacher characteristics and had also not taken all of the multiple-choice pretest questions, it was eliminated from the HLM analyses. Further, a handful of teachers from the other sites also failed to complete either the teacher questionnaire or parts of the assessment. Therefore, the final analytic sample for the HLM analyses was 234 teachers across 9 sites.

To ensure that there were no significant differences between the two groups on the assessments at pretest, we compared the means of the two groups. We excluded the trainer site from these analyses, because this site contained only treatment teachers. However, we did expect them to have slightly higher scores than other

Table 2
Descriptive Statistics for the 10 DMI Sites

Site			MC Pretest	MC Posttest	OE Pretest	OE Posttest
A	Comparison	Mean	.06 (.72)	.41 (.97)	–.11 (.82)	–.08 (.71)
	(*n* = 12)	(*SD*)				
	DMI	Mean	–.11 (.98)	.26 (.74)	.32 (.81)	.36 (.98)
	(*n* = 17)	(*SD*)				
B	Comparison	Mean	–.59 (1.17)	–.23 (1.14)	–.34 (.70)	–.13 (.67)
	(*n* = 10)	(*SD*)				
	DMI	Mean	–.29 (1.20)	–.10 (1.21)	–.36 (.87)	.07 (1.16)
	(*n* = 19)	(*SD*)				
C	Comparison	Mean	–.74 (1.07)	–.51(.90)	–.66 (.65)	–.79 (.59)
	(*n* = 16)	(*SD*)				
	DMI	Mean	–.26 (1.22)	.27 (.90)	–.24 (1.36)	–.08 (1.17)
	(*n* = 17)	(*SD*)				
D	Comparison	Mean	.22 (.71)	.19 (.61)	.34 (.84)	.14 (1.21)
	(*n* = 10)	(*SD*)				
	DMI	Mean	.04 (.35)	.46 (.27)	.09 (.49)	.13 (1.04)
	(*n* = 4)	(*SD*)				
E	Comparison	Mean	–.52 (1.11)	–.26 (1.16)	–.22 (.84)	–.15 (1.02)
	(*n* = 20)	(*SD*)				
	DMI	Mean	-07 (.76)	.34 (.78)	–.17 (.88)	.76 (.87)
	(*n* = 21)	(*SD*)				
F	Comparison	Mean	.39 (.86)	.86 (.56)	.77 (.70)	.33 (1.00)
	(*n* = 9)	(*SD*)				
	DMI	Mean	.39 (.91)	.72 (.80)	.31 (.65)	.78 (.89)
	(*n* = 6)	(*SD*)				
G	Comparison	Mean	–.72 (.97)	–.55 (.91)	–.58 (.87)	–.51 (.91)
	(*n* = 36)	(*SD*)				
	DMI	Mean	–.43 (.94)	.12 (.87)	–.60 (.79)	.09 (.84)
	(*n* = 30)	(*SD*)				
H	Comparison	Mean	.07 (1.09)	.11 (1.14)	.03 (1.20)	–.08 (1.16)
	(*n* = 8)	(*SD*)				

Table 2 (continued)
Descriptive Statistics for the 10 DMI Sites

	DMI (n = 8)	Mean (SD)	–.45 (.90)	.28 (.84)	–.24 (.92)	.07 (1.08)
I	Comparison (n = 8)	Mean (SD)	–.14 (.79)	24 (.68)	–52 (.44)	.01 (.46)
	DMI (n = 6)	Mean (SD)	–.17 (.99)	.47 (.58)	–.86 (.99)	.30 (.54)
J	DMI (n = 51)	Mean (SD)	.63 (.67)	.71 (.76)	.37 (.86)	.93 (.94)
Total	Comparison (n = 129)	Mean (SD)	–.37 (1.03)	–.17 (1.00)	–.27 (.89)	–.26 (.92)
	DMI (n = 179)	Mean (SD)	.02 (.98)	.37 (.87)	–.07 (.97)	.45 (1.02)

treatment teachers because of their interest in becoming trainers.[2] There were no statistically significant differences between the treatment group and the comparison group on either of the two pretest assessments. It is impossible to know whether the two groups of teachers were equivalent on other measures prior to the start of the training. It is also impossible to know whether the groups differed on their motivation to learn; however, given that participation in DMI is voluntary, it is likely that the two groups differed. Despite this, the similarity of the two groups' pretest scores suggests that the two groups had similar knowledge levels on the materials covered in the DMI training. The demographic information reported previously also partially supports a claim of group similarity.

Pretests and Posttests of Teacher Knowledge[3]

At each site, all participating teachers in both the DMI and comparison group took pretests and posttests designed to assess their MKT, with items designed to measure SCK, KCS, and KCT as characterized by Ball and colleagues (2008). The teachers took the pretest prior to participating in one module, Building a System of Tens (BST), and then the posttest after completing a second module, Making Meaning for Operations (MMO). Comparison teachers were administered the pretests and posttests at the same times DMI participants were administered the instruments.

[2] Because of the uniqueness of this site, we could have dropped it from all subsequent analyses. We elected not to do so after conducting analyses with and without the site and finding consistent results.

[3] For a more detailed explanation of the measures development, see Higgins, Bell, Wilson, McCoach, and Oh (2007).

The intervention was defined as the completion of this two-course series for several reasons. First, the BST–MMO sequence is the most frequently offered sequence at DMI sites. Thus, we wanted to understand the effects that are most likely to be occurring in sites that use DMI. Second, the BST–MMO sequence gives teachers enough time and exposure to the ideas in the curriculum that we might reasonably expect changes in knowledge to occur. Finally, the two modules are complementary and designed to engage teachers in some of the surprisingly complex but foundational ideas in the number and operations strand. These ideas are central to the K–8 mathematics identified in *Curriculum Focal Points* (National Council of Teachers of Mathematics, 2006). The modules tested were the original versions because those were the ones that were available at the time of the study. New editions of the modules were released in spring 2009.

Items for these tests were developed in two ways: Multiple-choice items were selected from the Mathematical Knowledge for Teaching (MKT) item bank (Ball, Hill, Rowan, & Schilling, 2002), and open-ended items were based on earlier items used by DMI project leaders during early field-testing of the materials. We describe each briefly.

Multiple-choice items. We created an instrument using items developed by the LMT project (Ball et al., 2002; Hill & Ball, 2004). For sample items see http://www.sii.soe.umich.edu/documents/released_items02.pdf. Several criteria were used to select appropriate items. First, we selected items identified by the LMT project as belonging to the number and operations strand. After checking this classification, we tagged all items that corresponded with mathematical topics covered in the two DMI modules being investigated (BST and MMO). For example, fractions were covered as a central topic in the DMI modules being studied, so any fractions items were eligible for inclusion in our measure.

For each item we noted whether it could be classified as addressing SCK or KCS.[4] Classifications were based on scoring tables provided by the LMT project to those completing an item training workshop (LMT Project, n.d.).[5] Specialized content knowledge (SCK) is mathematical knowledge that is specific to teaching. It goes beyond common content knowledge shared by other adults who use mathematics in their work. It includes analyzing students' methods (especially nonstandard methods), determining validity of a statement or solution process, selecting appropriate representations for teaching specific ideas or content, and so on. It goes beyond "knowing how" and can include "knowing why." Knowledge of content and students (KCS) is knowledge that teachers draw upon to understand student work and development. It can include knowledge of the typical errors students make,

[4] The SCK items were simply classified as content knowledge items, but we selected items that fit the definition of SCK more than CCK. At the time the items were selected more detailed classifications were not offered. When the classifications were reviewed, items were classified as SCK, CCK, neither, or both.

[5] Author Tracy Higgins attended the June 18, 2004, "instrument camp."

common strategies they are likely to employ, and ideas that can be confusing to them. The LMT items that we selected measured aspects of both content knowledge (i.e., SCK) and pedagogical content knowledge (i.e., KCS).

Once items with appropriate content had been identified, we then selected from them the items that proved most reliable in previous large-scale analyses reported by LMT, as well as items that represented a range of difficulty levels. We eliminated some questions that would lead to overrepresentation of some mathematical content and focused on questions that required analysis of student work, instructional tasks, representation tools, and content. When we had a defensible, representative set of items, the DMI advisory board—composed of mathematicians, DMI developers, professional development leaders, mathematics educators, and education researchers—reviewed the items for content, relevance, and face validity.[6] The final form included 13 SCK items (which included items focused on analyses of instructional representations and problems, evaluation of the validity of nonstandard methods, and assessment of the mathematical validity of student work) and 7 KCS items (which included items focused on analyses of student errors, assessments of student explanations, and evaluation of potential problems for assessing student understanding).

Open-ended items. We also developed a set of open-ended items. The items were based on existing instruments originally created by DMI authors. The instruments had evolved over several years and had been used as embedded assessments in some small-scale DMI projects. As embedded assessments in the DMI curriculum, we presumed that these items might be even more closely aligned with the DMI "treatment" than the LMT items. In contrast to the multiple-choice items, which only measured SCK and KCS, the open-ended items also measured knowledge of content and teaching (KCT). KCT is knowledge of how mathematics and teaching are brought together as teachers select examples that have specific instructional and mathematical features designed to support deeper student understanding, sequence mathematical examples toward some instructional goal, and assess various pedagogical approaches for their mathematical affordances.

The existing items were examined against the relevant BST and MMO learning goals in the domains of SCK, KCS, and KCT. Modifications were made to improve coverage of content represented in DMI's learning goals and to include each operation as well as both whole numbers and rational numbers. See Figure 1 for an excerpt from the open-ended assessment focusing on core ideas in multidigit multiplication.

The open-ended assessment had four stems, the first of which focused on subtraction. The subtraction items involved explaining student work, assessing the validity of students' methods, and generating additional examples of likely student work. The items were designed mainly to assess SCK, although KCS was also assessed. After final versions of the items and scoring rubrics were complete, items and

[6]This process is explained in greater detail in Higgins et al. (2007).

Problem 2. Imagine you are a fourth-grade teacher. At the beginning of the year, several of your students are making the same error in multiplication problems:

$$24 \times 36 = 20 \times 30 + 4 \times 6 = 600 + 24 = 624$$
$$16 \times 18 = 10 \times 10 + 6 \times 8 = 100 + 48 = 148$$

a. Explain what the students might have been thinking. Why does this method give the wrong answer?

b. List up to four things these students seem to know about solving multiplication problems of this type.

c. As their teacher, what are two approaches you might take to help students understand why this method doesn't work? (Include examples to clarify what each approach entails).

Scoring: Items within each stem test for content knowledge, understanding of student ideas and representations, and knowledge that combines these to arrive at instructional decisions. Each item (c. above counts as two items) is scored on a scale from 0–2. Coding is analytic with the rubric specifying content required to score at each level. The conceptual underpinning of the coding system is based on content specific questions keyed to each item—the rubric is designed to produce different scores for responses that vary in how well they address them. The questions underlying scoring for a–c are as follows:

a. Does the respondent understand how the students were breaking down the numbers? Can the respondent link the students' approach to a viable method that also begins with the decomposition of both numbers into tens and ones? Can the respondent point to what is missing or why the method did not work?

b. Rubric includes a list of student ideas that teachers could build upon. Coding by counting the number of ideas expressed and translating this into a score on a 0–2 scale.

c. Pick out two best approaches: Does the approach provide opportunities to explore how multiplication works using tools, representations, contexts, or numbers that are meaningful to students? Are students engaged in assessing whether their original answers were reasonable? Are activities suggested that will support students in identifying what is missing from their earlier work and seeing why those missing parts are important?

Figure 1. Excerpt from the open-ended teacher assessment.

rubrics were checked to ensure they still focused on the intended constructs. The second stem (displayed in Figure 1) involved a common error students make in multidigit multiplication. Three items assessed teachers' ability to identify and explain students' errors (SCK), to assess aspects of understanding the student showed (SCK and KSC), and to provide two different instructional approaches that could help students understand why their method did not work (KCT). The third

stem assessed knowledge of the range of student solution strategies that might be used to solve a word problem involving dividing something among friends. Although worded as a single item, multiple responses were required and those responses were weighted to be worth three items. We believe this item measures KCS. The final stem assessed understanding of fractions using a word-problem context similar to that found in one of the multiple-choice items. Four items presented different examples of student work. In each case, the task was to explain what the child was thinking and evaluate the validity of his or her answer. In each case, partial credit could be based on CCK and/or SCK, but we believe that full credit showed evidence of both SCK and KCS. One advantage of the underlying conception of MKT is that it acknowledges the closely knit nature of these varied forms of knowledge, although this often means that it is not always easy to map a single item onto one aspect of MKT. This led to some complexity in scoring.

The items were reviewed by the DMI advisory board for conceptual and face validity. In response to feedback, items were modified; we went through 12 drafts of the open-ended items. The final form was piloted with three groups of teachers, totaling 53 subjects from several states and a variety of school systems. Teachers in two of these groups provided feedback on the items' face validity and their responses were used during the development of scoring rubrics. The final form was composed of four stems and included 14 items.

The scoring rubric went through a lengthy development process. We began with conjectures about the range of responses we would receive for each item. We then used pilot data to examine the range of actual responses and began comparing them based on what we saw as evidence of strong, moderate, or weak MKT. Near-final versions were reviewed for content validity by the DMI authors and a local group of Boston-area researchers and practitioners who are DMI "critical friends"; final versions were tested with pilot data and data from individuals who did not complete the full pre–post study.

We used the same items for both pre- and postassessment. Developing and equating parallel forms were beyond our project's means. Although familiarity with items can lead to inflated scores at second testing, the testing sessions were at least 1 month apart, and in most cases, 6 months apart. DMI and non-DMI participants at each site took pre- and postassessments during the same time interval so that the time between testing would be the same for both groups. To ensure that copies of the items were not distributed between testing, we had the facilitators collect all assessment forms, even if they were not completed, and return them to us. We also instructed participants to use only the paper on which the distributed assessments appeared and to refrain from making copies of any of the questions or discussing them until after the study was complete.

Open-ended item scoring. After a day-long training session and subsequent calibration work, each coder scored the items associated with one of the stems for all data (blind to whether she was coding pre- or posttest data). Each item in the open-ended assessment was coded by a single scorer. Approximately 11% of the

open-ended responses were then scored by an additional coder to assess the degree of inter-rater agreement. The mean inter-rater agreement across the 14 open-ended items was 80.8%, which indicates that the pre- and postmeasures of MKT exhibited acceptable stability and consistency across raters.

Multiple-choice item scoring. The multiple-choice items were scored as correct or incorrect. For the purposes of the HLM analyses, we used the first 13 questions on the multiple-choice pretest, because most of the teachers from Site 1 took a version of the multiple-choice pretest that did not contain questions 14 and 15. This was due to a logistical error that occurred sometime between the distribution and administration of the testing materials. However, the multiple-choice posttest used for the HLM analyses contained all 15 questions. In general, more items allows for greater precision in scores. Though this is often true, our evidence does not support this claim for the posttest.

The multiple-choice assessment contained 20 questions, which were scored dichotomously. Overall scores thus could range from 0 to 20. The internal consistency reliability estimates for the multiple-choice pre- and postassessments were .77 for the pretest and .79 for the posttest. The open-ended assessment contained 14 items that were scored with a rubric, and overall scores could range from 0 to 32. The internal consistency reliability estimate for the open-ended pretest was .76; at posttest, the internal consistency reliability estimate was .79. Table 3 contains information regarding the subscale reliabilities and average interitem correlations.[7]

Surveying Cross-Site Variability

We were interested in the effects of DMI across all sites. However, based on visits to many of the sites, we also were aware that there might be important cross-site variability. After consultation with DMI developers and users, we developed a survey of DMI facilitators to gather data about program structures and group dynamics using an electronic questionnaire administered through SurveyMonkey© (SurveyMonkey, n.d.). The questionnaire focused on specific information regarding facilitator experience and characteristics, group dynamics, study administration (e.g., composition of comparison group, ease of recruiting comparison group), and module structure (length, timing, features, etc.). At each site, the study administrator, DMI module facilitator(s), and any relevant district personnel were asked to complete the survey. One study administrator, 9 study administrator/ facilitators,

[7]The correlation between the pretest administration of the multiple-choice assessment and the posttest administration of the multiple-choice assessment was .77. The correlation between the pretest administration of the open-ended assessment and the posttest administration of the open-ended assessment was .67 (see the Appendix). Although dissatisfying as a result, this is as one would expect given the indistinct line between teachers' mathematical knowledge for teaching and their pedagogical content knowledge, a point we mentioned earlier and one we return to later in the paper. There are, of course, other possible explanations, including the number of grade levels tested (our measures might be more prone to error for some grade levels) or technical difficulties (for some pilot testing, items were not formatted appropriately or consistently).

Table 3
Internal Consistency Reliability (*Cronbach's Alpha*)

Test	No. of items	Mean inter-item correlation (IIC)	Standard deviation of IIC	Cronbach's Alpha
Multiple-choice pretest	18	.15	.095	.77
Multiple-choice post-test	20	.15	.089	.79
Open-ended pretest	14	.19	.126	.76
Open-ended posttest	14	.222	.122	.79

6 facilitators, and 4 district personnel completed the questionnaire. Two of the 22 individuals who were asked to participate did not respond.

For this analysis, we elected to focus on variation across facilitators' knowledge and skills as our level-2 construct. We presumed that teacher learning would be enhanced if facilitators were more skilled and knowledgeable about DMI. This assumption, as pointed out previously, resonates both with current best practice in professional development more generally and with DMI in particular. Ideally, of course, we would have independent measures of facilitator knowledge and skill, not unlike the measures we were using for the teachers. However, no such measures exist, so we searched for a responsible proxy.

Based on our interviews with site personnel, it was clear that facilitators regarded as excellent by their peers had a range of experiences working with the DMI materials. Given this breadth of experience, it seems likely that a facilitator would have seen the ways teachers engage with the material. Also, more experienced facilitators likely have worked with the materials longer and therefore have higher levels of relevant knowledge.[8] Having many kinds of experiences with DMI may also be a proxy for some of the less tangible aspects of being a good facilitator—seeing oneself as a learner, sharing DMI's instructional stance, and being committed to DMI as a vehicle for professional development. Thus, we decided on a *breadth of opportunities to learn* (OTL) variable for facilitators that would be a proxy for facilitator knowledge and skill.

On the site questionnaire, facilitators were asked to respond to several questions regarding their opportunities to learn DMI. By summing responses to nine questions that assessed facilitators' breadth of DMI experiences, we constructed the OTL variable. These questions asked whether the facilitator had attended the

[8]We assume that there is as much conceptual and empirical work to be done in mapping out what mathematical and pedagogical knowledge facilitators need, and that the knowledge that teachers need is not isomorphic to what facilitators need.

leadership institute at Mount Holyoke College, attended a leadership institute at another location, apprenticed to another facilitator, cofacilitated, taught at the Mount Holyoke leadership institute, participated in a study group of other facilitators, had written cases about their facilitating, been observed by others, or participated in other non-DMI leadership training. Scores on the OTL variable ranged from 2 to 9, with a mean of 5.3 and a standard deviation of 2.5.

All facilitators had cofacilitated, most (8 of 10) had been observed by another facilitator, had facilitated other DMI modules, had apprenticed to someone else, and had facilitated BST and MMO four or more times. Roughly half the facilitators went to the Mount Holyoke institute, had been in a facilitator study group, or participated in another leadership training program. Only 3 had taught at Mount Holyoke or written cases about their teaching experiences.

Data Analysis

To evaluate the effectiveness of the DMI training, we compared the gain scores of teachers who received DMI training to scores of comparison teachers who were not involved in the training, after controlling for pretest scores. To analyze the intervention's effectiveness, we conducted a series of multilevel regression analyses using HLM 6.06. We analyzed the impact of the professional development training on two separate dependent variables: multiple-choice test score and open-ended test score. In these analyses, level 1 was the teacher level; level 2 was the site level. Because teacher demographic data and some pretest data were not collected from one site (Site 1), that site was eliminated from the HLM analyses. We estimated all models using restricted maximum likelihood estimation, given the small level-2 sample size (N = 9). For the analyses of each dependent variable, we used a model-building approach (Raudenbush & Bryk, 2002). First, we estimated the unconditional model, which allowed us to look at the variability across sites in terms of gains. Then we estimated a level 1 model, which included two independent variables—treatment (coded 0 for comparison and 1 for treatment) and pretest score, which was grand mean centered—and the interaction between them. Given the small number of level 2 units, we allowed only the intercept to randomly vary across sites. Finally, we estimated a level 2 model, with the OTL variable as a predictor of the level 1 treatment slopes, resulting in the estimation of a cross-level interaction between facilitator's OTL and the effect of treatment on the predicted dependent variable score.

RESULTS

DMI Implementation Across the Sites

There was variability across sites in terms of how DMI was structured. All teachers were volunteers, but four sites paid teachers for participating. A typical variation occurred in the timing of the module offering. All sites reported covering the entire curriculum; however, the time between sessions varied. Sites that held

weekly meetings completed the curriculum in 7–9 weeks; sites that generally met every other week took 12–16 weeks to complete a module (see Table 4). This was true for both BST and MMO. One site used a summer format, which lasted just over 1 week, meeting intensively every day. And one site used a mixture of these two formats. The average length of time between the beginning of BST and the end of MMO was 21.1 weeks or approximately 5 months. There was a wide range however, with the longest time being 1 year and the shortest 10 days.

DMI authors consider the regular assignment of homework and written feedback from facilitators as crucial to participants' learning. As prescribed, all facilitators reported that they assigned homework to teachers and gave written feedback on those homework assignments weekly (Table 4). Facilitators also reported that they used all or most of the materials provided by the developers (videos, readings, exercises, etc.).

Table 4
Implementation of Specific DMI Features (*Percent of Sites*)

	BST	MMO
Format of session		
After school	80	89
Summer	20	11
Session length		
15–16 weeks	20	22
12–13 weeks	30	22
7–9 weeks	40	44
Intensive 1 week	10	11
Assigned homework		
No	0	0
Some sessions	20	22
All sessions	80	78
Gave written feedback		
No	0	0
Some sessions	20	33
All sessions	80	67

In addition to variation in the structures that shaped teachers' learning contexts (e.g., the number of times they met, how often they met, whether they were paid), there was variation in the curricula teachers used in their classrooms. Even though most sites were school districts, only one of the sites, by teacher report, limited itself to a single mathematics curriculum. The average number of curricula reported

by teachers at a single site was 3.7, with four different curricula being the modal site experience. At one regional site, teacher-participants used nine different curricula or publishers: *Investigations*, Addison-Wesley, Scott Foresman, *Everyday Mathematics*, Harcourt Brace, Saxon, Prentice Hall, *Connected Math Project*, and McGraw Hill.

There was very little variation in facilitator reports about session quality. All site facilitators reported high (7 of 10) or moderate (3 of 10) levels of teacher engagement, group synergy, and levels of learning. There were no sessions of BST or MMO that facilitators rated as having low or no engagement, synergy, or learning. Facilitators unanimously rated DMI practices and materials as highly likely to be used by teachers and highly relevant to the participants' practice. From facilitators' perspectives, these sessions went well, produced learning, and met participants' needs.

Facilitators' and teachers' survey responses suggest that DMI was implemented with a high degree of fidelity to the structures valued by the developers. There were variations in the structural features of implementation—frequency of meetings, span of time to learn both BST and MMO, and the schools' curricular contexts. But core aspects of the seminar—covering all the content, using homework, and providing written feedback—remained intact.

Multiple-Choice Assessment

First, we estimated an unconditional model to calculate the intraclass correlation (ICC), which provides a measure of the ratio of between-site variability to total variability. The between-site variability was .02, which was not statistically significantly different from 0 ($p > .50$); the within-site variability was 7.09. Thus, the ICC was approximately .003, indicating that only about .3% of the variability in the gains in the multiple-choice assessment was between sites. Given the negligible amount of between-site variability, we did not need to estimate a random effect for the intercept. Thus, multilevel modeling was unnecessary for the analyses of the multiple-choice gain scores. We continued our regression-based analyses in HLM for consistency of presentation. However, given the lack of random effects in the model, these results are equivalent to the results from a single-level regression model.

Our full level 1 model (1) included pretest scores (grand-mean centered), treatment group (coded 0 for comparison and 1 for treatment), and the pretest by treatment interaction.[9]

$$\begin{aligned} MCGain_{ij} = \beta_{0j} + \beta_{1j}(TRT)_{ij} &+ \beta_{2j}(MCPRE)_{ij} \\ &+ \beta_{3j}(MCPRE)_{ij} * (TRT)_{ij} \\ &+ \beta_{4j} PROFDEV_{ij} + r_{ij} \end{aligned} \quad (1)$$

However, the pretest-by-treatment interaction and the professional development variables were not statistically significant. Therefore, we eliminated these variables

[9]In all equations, *ij* refers to individual *i* within site *j*.

from our final level 1 model. The mean gain for teachers in the comparison group was .66 (γ_{00}), and this was statistically significantly different from 0 ($t = 2.58$, $p = .01$), indicating that, even in the absence of intervention, teachers' scores on the multiple-choice assessment increased from pretest to posttest. Both pretest score ($\gamma_{20} = -.24, p < .01$) and treatment group ($\gamma_{10} = .97, p = .01$) were statistically significant predictors of teachers' gains on the multiple-choice assessment. For teachers whose pretest scores were above the mean, every point increase in pretest scores resulted in an expected decrease of .24 in teachers' gain scores. In other words, teachers with the highest pretest scores were expected to have the lowest gain scores. The treatment group outperformed the comparison group by almost 1 point, indicating that the gain scores of the treatment group were about 1 point higher than the gain scores of the comparison group. This translates to a Cohen's *d* of .36, indicating that the treatment group moderately outgained the comparison group from pretest to posttest. Adding pretest scores to the model explained 10.6% of the variance in teacher gains. Adding treatment to the model explained an additional 3% of the variance in teacher gains, in addition to that which was explained by pretest.

Our level 2 model (2) included a cross-level interaction between facilitator OTL and the effect of treatment on teacher gains.

$$MCGain_{ij} = \beta_{0j} + \beta_{1j}(TRT)_{ij} + \beta_{2j}(MCPRE)_{ij} + \beta_{3j}(OTL)_j * (TRT)_{ij} + r_{ij} \quad (2)$$

Facilitator OTL was not a statistically significant moderator of the effect of treatment on teacher gains. For exploratory purposes, we also fit models using several other facilitator variables (attendance at the Mount Holyoke institute, participation in a study group, years of facilitator experience, and facilitation of other DMI modules). None of these facilitator-level variables moderated the effect of treatment on teacher gains. Thus, our final level 1 model (3) appears as our best model in Table 5.

$$MCGain_{ij} = \beta_{0j} + \beta_{1j}(TRT)_{ij} + \beta_{2j}(MCPRE)_{ij} + r_{ij} \quad (3)$$

Open-Ended Assessment

Next, we ran a series of multilevel models to predict teacher gains on the open-ended assessment. First, we estimated an unconditional model to calculate the intraclass correlation. The between-site variability was 1.77, which was statistically significantly different from 0 ($p < .01$); the within site variability was 21.90. Thus, the ICC was .075, indicating that approximately 7.5% of the variability in the gains on the open-ended assessment was between sites. We used multilevel modeling to model the between-site variability in the gains on the open-ended assessment.

Next, we fit a level 1 model that included pretest scores, amount of professional development training teachers reported having received, treatment group, and a

Table 5
Final Regression Model for Multiple-Choice Gain Score

Fixed effects	Coefficient (*SE*)	*t* (*df*)	*p*
Model for intercept (β_0)			
Intercept (γ_{00})	.66 (.26)	2.58 (231)	.011
Model for TRT slopes (β_1)			
Intercept (γ_{10})	.97 (0.34)	2.87 (231)	.005
Model for MC PRETEST slopes (β_2)			
Intercept (γ_{20})	–.24 (.04)	–5.91 (231)	<.001
Residual variance ($\sigma^2 = 6.16$)			

treatment-by-pretest interaction effect as predictors of teachers' gains (4).

$$OEGain_{ij} = \beta_{0j} + \beta_{1j}(TRT)_{ij} + \beta_{2j}(MCPRE)_{ij} + \beta_{3j}(MCPRE)_{ij} * (TRT)_{ij} + \beta_{4j} PROFDEV_{ij} + u_{0j} + r_{ij} \quad (4)$$

However, the pretest-by-treatment interaction was not statistically significant. Therefore, we eliminated this variable from our final level 1 model. The mean gain for teachers in the comparison group was .17 (γ_{00}); this was not statistically significantly different from 0 ($t = .30$, $p = .77$). Both pretest score ($\gamma_{20} = –.28$, $p < .001$) and treatment group ($\gamma_{10} = 3.42$, $p < .001$) were statistically significant predictors of teachers' gains on the open-ended assessment. For teachers whose pretest scores were above the mean, every point increase in pretest scores resulted in an expected decrease of .28 in teachers' gain scores. Again, teachers with the highest open-ended pretest scores were expected to have the lowest gain scores. The coefficient for mathematics professional development (γ_{30}) was .38, indicating that after controlling for treatment group and pretest scores, teachers with greater professional development training experienced greater gains on the open-ended assessment. After controlling for professional development experience and pretest scores, the treatment group outperformed the comparison group by over 3.4 points, indicating that the gain scores of the treatment group were almost 3.5 points higher than the gain scores of the comparison group. This translates to a Cohen's *d* of almost .70, indicating that the treatment group substantially outgained the

comparison group from pretest to posttest. Adding pretest scores to the model explained 8.4% of the within-site variance in teacher gains. Adding teachers' prior professional development experience to the model explained an additional 3% of the within-site variance in teacher gains, in addition to that which was explained by pretest. Finally, adding treatment group to the model explained an additional 11% of the within-site variance in teacher gains, in addition to that which was explained by pretest and professional development experience.

Finally, we added facilitator's OTL (grand-mean centered) as a moderator of the effect of treatment on teachers' gains on the open-ended assessment (5).

$$OEGain_{ij} = \beta_{0j} + \beta_{1j} + \beta_{1j}(TRT)_{ij} + \beta_{2j}(MCPRE)_{ij} + \beta_{3j}(OTL)_{ij} * (TRT)_{ij} + \beta_{4j}PROFDEV_{ij} + u_{0j} + r_{ij} \quad (5)$$

This effect ($\gamma_{11} = .46$) was statistically significant ($p = .03$), indicating that the higher the facilitators' OTL score, the greater the difference between the gain scores of the teachers in that site. The coefficients for the final model appear in Table 6 but are quite similar in magnitude and interpretation to those reported previously for the level 1 model. After controlling for all the variables in the model, the mean gain in the comparison group was still not statistically significantly different from 0 ($\gamma_{00} = .57, p = .38$). The treatment group still exhibited an average advantage of

Table 6
Final HLM for Open-Ended Assessment Gain Scores

Fixed effects	Coefficient (*SE*)	*t* (*df*)	*p*
Model for intercept (β_0)			
Intercept (γ_{00})	–.57 (.61)	–.9(8)	.38
Model for TRT slopes (β_1)			
Intercept (γ_{10})	3.38 (.59)	5.76 (229)	<.001
OTL (γ_{11})	.46 (.21)	2.14 (229)	.03
Model for PRETEST slopes (β_2)			
Intercept (γ_{20})	–.28 (.05)	–5.40 (229)	<.001
Model for Math PD slopes (β_3)			
Intercept (γ_{30})	.32 (.19)	–1.69 (229)	.09
Random effects (Variance components)	Variance	χ^2(*df*)	*p*
Between site var. in intercepts (τ_{00})	.47	11.57(8)	.17
Var. within sites (σ^2)	17.45		

almost 3.4 points over the comparison group. However, given the statistically significant cross-level interaction between OTL and treatment, this advantage was even stronger in sites where the facilitator had a high OTL score, and the advantage was smaller in sites where the facilitator had a below-average OTL score. After controlling for all variables in the model, the effects of pretest score (–.28) and mathematics professional development (.32) were also of a similar magnitude as the level 1 model. However, in this final model, the effects of professional development were no longer statistically significant ($p = .09$). Adding OTL as a predictor of the treatment effect reduced the between-site variance in the intercept from .94 to .47, a decrease of 50%. Furthermore, in this final model, the between-site variability in gain scores was no longer statistically significantly different from 0 ($\boldsymbol{\tau}_{00} = 11.57$, $p = .17$).

Comparison of Multiple-Choice and Open-Ended Assessments

Although the DMI group did exhibit greater gains than the comparison group on both assessments, the magnitude of this difference was much larger on the open-ended assessment. This is not surprising: the multiple-choice items were chosen because they covered the same mathematical content as the DMI modules (subtraction, addition, division, multiplication, whole numbers, fractions, some work with decimals, etc.); the open-ended items were designed to assess DMI-specific learning goals, were aligned well with the big ideas explored in the context of the two-course seminar, and because they were developed from embedded assessments in the curriculum, were probably framed in ways more familiar to participants. In addition, although facilitator's OTL appeared to have an impact on the magnitude of the differences between the treatment and comparison groups on the open-ended assessment, this was not the case for the multiple-choice assessment. Although the measures differed in how closely aligned they were with DMI content, they were both designed to assess MKT and were highly correlated with each other (see the Appendix).

Where Is the Learning?

In order to better understand the nature of the learning captured in the multiple-choice and open-ended instruments, we conducted an item-by-item difference analysis of the two instruments (using *t* tests). This analysis points to items on which the treatment teachers learned more than the comparison group. Because the *t* tests are a liberal test of the relationships, we do not focus on statistical significance and instead report the general patterns of the content on which DMI participants seemed to be making gains. We report these in order to generate hypotheses others might test at a larger scale. On the multiple-choice items, teachers seemed to gain the most on two items, one measuring SCK, the other KCS. The first focused on place value knowledge (recognizing an error in interpretation) and the second engaged teachers in reasoning about student work involving fractions in a word problem context that was similar to one of the open-ended items. On the open-ended

instrument, five items showed learning gains by the DMI group. These included items concerned principally with SCK and items that required teachers to draw on KCS and KCT. These items asked teachers to analyze student work in a two-digit subtraction context, generate additional methods students might use in such a subtraction context, generate potential solution strategies students might employ in a word-problem context, describe next steps to help students understand why a solution strategy they used did not work, and explain the approaches used by students to solve a word problem involving fractions.

These exploratory analyses suggest to us the following hypothesis: Teachers who participate in DMI exhibit learning on questions that cover SCK, KCS, and KCT, all three domains of MKT that are directly addressed in the BST and MMO seminars. Teachers exhibit less learning on questions that assess MKT more generally within the K–8 number and operations topic area and that may not be as closely aligned with the problems explored in the DMI modules. It may also be that the multiple-choice items evoked more guessing strategies than the open-ended items, that our constructed multiple-choice forms were too brief to detect subtle distinctions that occurred due to DMI learning, that the open-ended items that measure KCT (not present in the multiple-choice items) were sensitive to something important in DMI seminars, or that the task of generating responses (rather than selecting among responses) is more sensitive to the type of learning that occurs in DMI than is responding to multiple-choice questions. Further research should investigate these possibilities.

DISCUSSION

What Did Teachers Learn?

One goal of this inquiry was to understand better whether teachers gained MKT through participating in DMI and, if so, what kinds of MKT they acquired. Two issues warrant further discussion. The first issue entails understanding the relationship of MKT with other valued aspects of education. The second issue involves considering how we might explain these increases in knowledge.

The claims we can make about what the improvement in teachers' MKT might mean for other valued aspects of education are limited. The open-ended measure of teaching knowledge has not yet been associated with student learning or teaching practices. However, it is worth noting that the MKT items have considerable validity evidence associated with them. When administered in the three original forms, the items were found to be related to student achievement (Hill et al., 2005) and to teaching practices (Blunk, 2007; Hill et al., 2008).[10] We used a modified form in

[10]Test forms are specified by the mode in which they are administered (paper and pencil, online, etc.), the order and format of the items, and the items themselves. Forms belonging to the same test are often equated, resulting in forms that are of similar difficulty and content coverage. The form we used was very similar to one used by the LMT project, but was not equated with the form from which it was derived.

order to ensure overlap with the content of BST and MMO; however, our form is similar enough to the original three that there is some reason to suspect changes on our form may be related to changes in other measures of teaching and learning.[11] Future work should examine these relationships more directly. Our current work looks at only one potential area of learning, namely, teacher knowledge. There may (or may not) be effects in other areas such as teaching practices, teacher beliefs, or student achievement.

We have some conjectures about why DMI participants demonstrated more knowledge than the comparison group. These results may have arisen because there was a larger percentage of secondary teachers and administrators, specialists, or nonclassroom teachers in the DMI group, as compared to the non-DMI group. This could positively bias the results because modules on number and operations could be viewed as a somewhat artificial treatment. We ran our analyses with and without this group in the sample and the pattern of results was unchanged. Further, anecdotal evidence from the sites suggests that secondary mathematics teachers were grateful for the opportunity to engage with such content because they regularly must address students' fundamental mathematical understandings if they are to successfully teach secondary content. If the DMI treatment were artificial in some way, we would not expect secondary teachers to embrace the content so eagerly. We would also expect significant knowledge differences between these subgroups and the rest of the teachers to appear on pretest measures. No such differences were found.

A second alternative concerns the counterfactual. The DMI participants may have improved their MKT without the intervention of DMI. As is the case with all counterfactuals, there is no way to test this. It is simply unknowable. However, the comparison group was in the same district, experiencing a similar district context for learning. We have a sense of the learning that took place for the comparison group. Given the moderate to substantial effect sizes we see for DMI teachers, it is plausible that the knowledge growth we see in DMI participants is caused by DMI. Yet, given the study's quasi-experimental design, which allowed participants to select into the treatment condition, it could be argued that the DMI group might over-represent teachers who are more motivated to learn than their colleagues are. Indicators of this bias could include elevated pretest knowledge scores and/or more hours of previous professional development at the onset of the study in the treatment group than in the comparison group. We saw little evidence of any kind of systematic bias across sites and have statistically controlled for pretest knowledge scores and number of hours of previous professional development in our analyses. Controlling for these may remove some of the influences that could jeopardize the internal validity of the study.

A final alternative explanation suggests that there may be something about DMI's content, character, or structure that makes a difference. DMI is well specified at the teacher and facilitator level. It is a coherent, long-term, and narrowly focused

[11]Confirmatory factor analyses (Higgins et al., 2007) suggest that, at a minimum, our form of the items has some internal validity.

approach on particular topics in K–8 mathematics. It requires teachers to move back and forth between seminars and their own classrooms, receiving regular, written feedback from facilitators. Finally, DMI encourages teachers to learn from their practice. Changes in teachers' scores on our assessments may be the result of the learning that occurred in these seminars.

The point about classroom practice is worth emphasizing: As Ball and Cohen (1999) argue, teachers' learning should be embedded in practice. After all, they spend most of their professional time working with students, having experiences that are relevant to their learning and development. But they are also most often alone during those experiences, and professional development opportunities are seen as something that takes place apart from their practice, which means that we may be missing a substantial opportunity to leverage teacher learning in and from their daily work. DMI is quite different in this regard, for it encourages teachers to take their nascent SCK, KCS, and KCT into their classrooms and try things out. Repeatedly, teachers told us of their revelations—both in seminars and in their own schools—as they drew on their growing knowledge of and enthusiasm for mathematics and teaching mathematics in their classrooms. This anecdotal evidence aligns with results from S. Cohen's (2004) yearlong study of changes in teachers' thinking and practices over the course of their participation in DMI seminars.

How Did Sites Structure DMI Implementation?

A second goal of this inquiry was to understand how sites structured their implementation of DMI. As described previously, there was variation in how sites offered the modules in terms of the frequency of sessions and length of time across BST and MMO. There was little variation in how the sessions went or carried out critical features of the seminars. Facilitators thought the sessions went well, and participants completed homework, received written feedback on homework, and used the entire curriculum.

Given the persuasive literature on the complications of policy implementation (e.g., D. K. Cohen, Moffitt, & Goldin, 2007; Honig & Hatch, 2004), we expected to see more structural variation in DMI implementation. There are a number of potential reasons for the level of variation we saw. It is possible that our site selection (which focused on sites that had been using DMI for more than 1 year) could bias implementation variation. It is also possible that because these sites knew they were participating in a research study, they either reported more fidelity to the program or they actually enacted the program with higher fidelity. It could also be that because of the integration inherent in DMI materials, facilitators were less likely to omit entire program components (such as homework or videos of practice). Perhaps the variation in implementation comes in the form of the facilitator changing the balance among components, for example, decreasing the number of homework assignments or showing fewer videos. If this happened, our implementation measures were not fine-grained enough to have detected such distinctions. Finally, it is possible that because most of the facilitators had experienced BST and MMO as learners themselves (and all of them regarded that learning experience

highly), the facilitators were hesitant to create program mutations. We cannot investigate these possibilities herein, but these explanations nominate aspects of implementation that others might consider building into Borko's (2004) Phase 2 research.

Given the level of variation we did detect, evidence from the open-ended items suggests that there was something different happening at each site. The results from the open-ended items, which are very closely aligned with the topics and the format of tasks in DMI training, suggest that the broader a facilitator's opportunities to learn about DMI, the more teacher learning he or she was able to facilitate. We hypothesize this increased effectiveness comes from the development of deeper knowledge of the content of DMI, including knowledge of mathematics, knowledge of students' learning of mathematics, knowledge of how to teach adults, and familiarity with the common mistakes and misconceptions teachers have when learning to teach children mathematics. More extensive knowledge of DMI also may have taught facilitators which ideas, exercises, or processes are most powerful for teachers. If teachers develop pedagogical content knowledge over time, so, too, professional development leaders might develop their own kind of pedagogical content knowledge related to facilitating teacher learning.

The substantial influence of facilitator OTL on open-ended learning gains raises questions about why we do not see this same mediation of teacher learning on the LMT items. Although we cannot investigate this thoroughly with the present data, we hypothesize that this is the result of the looser relationship between DMI content and the LMT measures. The LMT measures cover number and operations content broadly conceived, but the topical emphases of the items and the specific student approaches assessed in the items are not a perfect match to the BST and MMO curriculum. There are two other number and operations modules of DMI. This relationship between facilitator OTL and the LMT might differ if all the DMI modules were a part of the DMI training. The tests we use matter to the judgments we make about learning. It is possible that there was enough lack of overlap that we simply could not detect the role of the facilitator's OTL. Of course, the study design and size does not allow us to test the conjectures described here, and so we leave it for future research to mount the large-scale, resource-rich efforts necessary to test such hypotheses.

IMPLICATIONS

As previously discussed, there is a need to understand what it takes to move from locally designed and locally administered assessments of professional development to more robust, defensible measures that connect to existing understandings of teachers' mathematical knowledge for teaching. We now turn to a description of the challenges researchers face as they set about doing this work.

These types of studies are expensive—financially and politically. Locating sites, negotiating access, collecting the data, and creating databases are resource-intensive tasks, fraught with potential pitfalls. We paid participants, site administrators,

and raters. We asked sites to work with us in novel ways that infringed on their "real" jobs. This required us to use political capital that DMI had accumulated through its own hard work in respecting and responding to teachers and professional development leaders in specific sites. Although study administrators at the sites were gracious and helpful, the study was still an inconvenience. Thus, in calculating the costs of these types of studies, one must figure in the cost of using the goodwill that exists among site personnel or the need to build such capital and goodwill in order to sustain data collection.

These types of studies hinge on a researcher's ability to select an appropriate program for what Borko (2004) termed Phase 2 study. Even though many programs may be implemented in multiple sites and therefore could be candidates for such inquiry, researchers must be selective so as to avoid spending a great deal of time, money, and political goodwill to show that learning is weak because implementation is weak. It is important to understand why implementation is weak and the factors that shape implementation; however, if the field hopes to learn more about what participants learn from the program, researchers must make choices about which programs to study. Reflecting on the implementation of DMI, a critical component for researchers to assess is a program's ability to strike a productive balance between what needs to be "tight" in implementation and what can be "loose." DMI is loose about the ways the course can be structured. Districts can adapt the sessions and materials to the frequency and timing of their local contexts. The materials can and should be used with the existing curricula with which teachers work. However, DMI is tight about the integration of curricular materials. Teacher materials are tightly connected to facilitator materials and both pay careful attention to key learning opportunities for teachers. Materials are extensive and exhaustive. In order to learn something about teacher learning in Phase 2 research studies, investigators must be able to assess the degree to which a program has achieved an appropriate balance of tight and loose so that there is teacher learning to document.

Finally, this type of work necessarily builds on previous work. The original Teaching to the Big Ideas project that launched the development of the DMI modules began more than 10 years ago. In those intervening 10 years, developers learned more about the requirements of effective DMI training, incorporating what they learned in subsequent drafts of modules. They developed long-standing relationships with educators around the country. As the DMI developers learned more about how and what teachers learned from the seminars, they developed formative assessments that we eventually used as the starting point for the development of our open-ended items. In short, those 10 years of work made this study possible. This is a point that was driven home by the National Research Council's (2003) call for strategic research partnerships that would create stable platforms for research that could both grow from practical concerns and inform educational practice.

The nature of this work demands much from professional developers, researchers, and funding sources alike. It requires long-term commitment to the production and refinement of professional development materials. It requires stability in both

funding and vision. And it also requires strong relationships with practitioners. These demands suggest that calls for the documentation of teacher learning in professional development settings will not be answered in the short term. It also suggests that the field needs to think more carefully about how we train researchers to do this work; how we report the complexities of the work so others can learn from it; and how we support the development of long-term collaborations on funding and tenure cycles, which tend to be quite short.

CONCLUSION

The study reported here is a modest one, and its purpose was equally modest: to explore the relationships among teacher learning, facilitator experiences, and program features in one nationally disseminated professional development project. Although the purpose of the study was modest, it provides an important first example of the ways in which these fundamental supports for teacher learning interact to support teachers' development of MKT. We found that DMI participants demonstrated significantly increased MKT as assessed by both the multiple-choice assessment and the open-ended assessment, although the strength of that relationship was stronger for the open-ended assessment that was developed from a DMI-created tool. We also found that teachers who worked with facilitators who had broader opportunities to learn DMI did better on the open-ended assessment. The learning we documented and its relationship to facilitator experiences occurred in the context of a program that adheres to the consensus view of high-quality professional development. DMI immerses teachers in subject-specific, practice-based, long-term learning opportunities.

If the country needs large numbers of teachers to develop the MKT necessary for teaching students rich mathematics, we must have programs that can function effectively on a large scale. This study documents that a professional development program can produce evidence of changes in teachers' MKT in multiple sites with multiple facilitators. The effort also documents the substantive and structural aspects of one program that varied when it was scaled across the country. Together, these contributions begin to nominate critical aspects of program features, facilitators, and contexts that may support teacher learning and that are important to preserve and study when investigating the fidelity of "scaling up efforts." There is still, however, much work to be done on both DMI and other professional development programs. Equally important, both for research on professional development and teacher learning more generally, is the need for sustained work on and investment in the development of associated measures and instrumentation. Taken together, the study responds in important, if limited, ways to calls for better empirical evidence to ground claims about high-quality professional development.

REFERENCES

American Federation of Teachers. (2002). *Principles for professional development: AFT's guidelines for creating professional development programs that make a difference*. Washington, DC: Author.

Ball, D. L. (1989). Knowledge and reasoning in mathematical pedagogy: Examining what prospective teachers bring to teacher education (Doctoral dissertation, Michigan State University, 1988). *Dissertation Abstracts International, 50* (02), 416A.

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

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3–32). San Francisco: Jossey-Bass.

Ball, D. L., Hill, H. C., Rowan, B., & Schilling, S. (2002). *Measuring teachers' content knowledge for teaching: Elementary mathematics release items 2002*. Ann Arbor, MI: Study of Instructional Improvement. Retrieved June 16, 2004, from http://www.sii.soe.umich.edu/documents/released_items02.pdf

Ball, D. L., & Rowan, B. (2004). Introduction: Measuring instruction. *Elementary School Journal, 105*, 3–10.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*, 389–407.

Blunk, M. L. (2007, April). *The QMI: Results from validation and scale-building*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3–15.

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005). *Studying teacher education: The report of the AERA Panel on Research and Teacher Education*. Mahwah, NJ: Erlbaum.

Cohen, D. K., Moffitt, S. L., & Goldin, S. (2007). Policy and practice: The dilemma. *American Journal of Education, 113*, 515–548.

Cohen, S. (2004). *Teachers' professional development and the elementary mathematics classroom: Bringing understanding to light*. Mahwah, NJ: Erlbaum.

Desimone, L., Porter, A. C., Birman, B. F., Garet, M. S., & Yoon, K. S. (2002). How do district management and implementation strategies relate to the quality of the professional development that districts provide to teachers? *Teachers College Record, 104*, 1265–1312.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*, 81–112.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: U.S. Department of Education.

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

Goldhaber, D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*, 129–145.

Guyton, E., & Farokhi, E. (1987). Relationships among academic performance, basic skills, subject matter knowledge, and teaching skills of teacher education graduates. *Journal of Teacher Education, 48*, 37–42.

Higgins, T., Bell, C. A., Wilson, S. M., McCoach, D. B., & Oh, Y. (2007, March). Measuring the impact of professional development on mathematical knowledge for teaching number and operations to elementary students. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Atlanta, GA.

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

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

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430–511.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. Amer*ican Education 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.

Honig, M. I., & Hatch, T. C. (2004). Crafting coherence: How schools strategically manage multiple, external demands. *Educational Researcher, 33*(8), 16–30.

Kennedy, M. M. (1999). The problem of evidence in teacher education. In R. A. Roth (Ed.), *The role of the university in the preparation of teachers* (pp. 87–107). London: Falmer Press.

Kennedy, M. M., Ball, D. L., & McDiarmid, G. W. (1993). *A study package for examining and tracking changes in teachers' knowledge* (Technical Series 93-1). East Lansing, MI: The National Center for Research on Teacher Education.

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

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

National Academy of Education. (2009). *Teacher quality* (S. M. Wilson, Ed.). Washington, DC: Author.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: Author.

National Research Council. (2003). Strat*egic education research partnership* (M. S. Donovan, A. K. Wigdor, & C. E. Snow, Eds.). Washington, DC: The National Academies Press.

Porter, A. C., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 259–297). Washington, DC: American Educational Research Association.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Schifter, D., Bastable, V., Russell, S. J. (with Cohen, S., Lester, J. B., & Yaffee, L.). (1999). *Developing mathematical ideas: Number and operations, part 1. Building a system of tens: Casebook*. Parsippany, NJ: Dale Seymour.

Schifter, D., Bastable, V., Russell, S. J. (with Yaffee, L., Lester, J. B., & Cohen, S.). (1999). *Developing mathematical ideas: Number and operations, part 2. Making meaning for operations: Casebook*. Parsippany, NJ: Dale Seymour.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*, 454–499.

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

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1–22.

Stein, M. K., Silver, E. A., & Smith, M. S. (1998). Mathematics reform and teacher development: A community of practice perspective. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp.17–52). Mahwah, NJ: Erlbaum.

Stein, M. K., Smith, M. S., & Silver, E. A. (1999). The development of professional developers: Learning to assist teachers in new settings in new ways. *Harvard Educational Review, 69*, 237–269.

SurveyMonkey. (n.d.). [Web-based survey tool]. Palo Alto, CA: SurveyMonkey. Retrieved April 2, 2005, from http://www.surveymonkey.com

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education, 24*, 173–209.

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations*. Seattle, WA: University of Washington Center for the Study of Teaching and Policy.

Wilson, S. M., Shulman, L. S., & Richert, A. E. (1987). "150 different ways" of knowing: Representations of knowledge in teaching. In J. Calderhead (Ed.), *Exploring teachers' thinking* (pp. 104–124). London: Cassell Educational Limited.

Wilson, S. M., & Youngs, P. (2005). Research on accountability processes in teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA Panel on Research and Teacher Education* (pp. 591–644). Mahwah, NJ: Erlbaum.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved April 14, 2010, from http://ies.ed.gov/ncee/edlabs

Authors

Courtney A. Bell, Educational Testing Service, Research and Development, Rosedale Road, MS 04-R, Princeton, NJ 08541; cbell@ets.org

Suzanne Wilson, Michigan State University, Department of Teacher Education, 209 Erickson Hall, East Lansing, MI 48824; swilson@msu.edu

Traci Higgins, TERC, 2067 Massachusetts Avenue, Cambridge, MA 02140; traci_higgins@terc.edu

D. Betsy McCoach, University of Connecticut, Department of Educational Psychology, 249 Glenbrook Road Unit 2067, Storrs, CT 06269; betsy.mccoach@uconn.edu

Accepted June 23, 2010

APPENDIX

Correlations Among Multiple-Choice and Open-Ended Subscales at Pretest and Posttest

	MC Pre	MC Post	OE Pre	OE Post
MC Pre	1			
MC Post	.772	1		
OE Pre	.662	.575	1	
OE Post	.627	.683	.673	1

Note. All correlations are significant at the 0.01 level (2-tailed).