

Measuring the effects of professional development:

The case of developing mathematical ideas

by

Courtney Bell, University of Connecticut

Suzanne Wilson, Michigan State University

Traci Higgins, TERC

Abstract

The research we report examines the impact of a nationally disseminated professional development program, Developing Mathematical Ideas (DMI), on teacher's specialized knowledge for teaching mathematics. DMI participants were compared with colleagues from similar schools in the same region. Teacher knowledge was measured with two instruments: multiple choice items developed by the Study of Instructional Improvement and open-ended items developed primarily from assessments previously used by DMI. After controlling for pretest scores on both assessments, a hierarchical linear model suggested there were statistically significant differences between the two groups; the DMI group outperformed the comparison group on both assessments. Gains in teachers' scores were related to the degree of facilitator experience with DMI. Limitations of the study and challenges associated with documenting the relationships among teacher learning, facilitator experience, and professional development program features are discussed.

Introduction

It is now apparent that most of our efforts at school reform will come to nothing unless teachers are up to the task. Standards-based reform may be the lever that sets in motion the improvements the United States has sought in schools for decades. But common sense, parental experience, and the research literature are clear: The most successful school innovations rest on the time, talent, and skill of teachers. (The College Board, 2006, p. 9)

There seems little disagreement among policymakers, researchers, educators, administrators, or reformers that teachers are the critical component in improving U.S. education. Further, there is apparently universal agreement that high quality, on-going professional development for teachers is equally necessary:

The nation can adopt rigorous standards, set forth a visionary scenario, compile the best research about how students learn, change textbooks and assessments, promote teaching strategies that have been successful with a wide range of students, and change all the other elements involved in systemic reform – but without professional development, school reform and improved achievement for all students will not happen. (American Federation of Teachers, 2002, p. 22)

But traditional “in-service” will not do, for in the last 15-20 years, there have been countless calls for the improvement of such professional development. Even more recently, there have also been calls to collect data on the effectiveness of professional development programs. The research reported here grew out of the interest of the creators of one such professional development project – Developing Mathematical Ideas (DMI) – in assessing what teachers learned from participating in DMI.

This research is “Phase 2” professional development research (Borko, 2004). We document teacher learning in a single program across multiple sites with multiple facilitators. The goal of the research is to investigate teacher learning in DMI seminars taught by facilitators with different degrees of experience that were already out in the field doing this type of work.¹ This type of analysis facilitates new insights into the relationships among teacher learning, facilitator experience and professional development program features.

We begin by introducing DMI, then briefly review the extant literature on professional development and related attempts to measure teachers’ mathematical knowledge and teacher learning. We then describe the questions upon which this research is based, the methods of our inquiry, and our results. In conclusion, we explore some important factors that shape both research on teacher learning in professional development and the phenomenon of teacher learning more generally.

Developing Mathematical Ideas

Developing Mathematical Ideas is a professional development seminar designed for practicing K-8 teachers of mathematics (S. Cohen, 2004; Davenport & Morse, 2001; Remillard & Geist, 2002). Developed by Deborah Schifter, Virginia Bastable and Susan Jo Russell, the program includes a series of modules that have been pilot- and field-tested, and then published by Dale Seymour Publications (e.g., Schifter, Bastable, Russell, Cohen, Lester, & Yaffee 1999; Schifter, Bastable, Russell, Yaffee, Lester, & Cohen, 1999). Each module is designed as a series of eight working sessions, lasting three hours each. Materials are provided to support facilitators in developing a learning context within the seminar that models the type of learning context envisioned for the

elementary and middle school classroom. The role of the facilitator is key, in that much of the information about pedagogy is conveyed implicitly through example, while participants work through activities in which mathematical content and children's ideas are the central focus (Cohen, 2004). Currently, there are seven modules that cover topics ranging from numbers and operations to algebraic thinking. Each module is designed to help teachers develop their own mathematical understanding, their understanding of typical student ideas associated with the mathematics in question, and strategies teachers can use to develop students' mathematical ideas.

During Phase I of the development of DMI, the program developers had tight control over the ways in which facilitators were prepared, the materials used, and the timing and content of the sessions. As DMI has been disseminated, there have been considerable adaptations, and thus it is challenging to discuss how DMI is implemented across various sites. Here we offer some descriptors that apply no matter the adaptation. Teachers regularly attend sessions that have been laid out carefully in professional development curricular materials. The sessions can be offered over a half year, or in more intensive week-long sessions. Each session is organized around cases of actual classroom experiences involving students and their thinking about mathematics. The cases are accompanied by focal questions designed to guide productive discussions, video clips, mathematics problems designed to engage adult learners, and relevant student work. Each session involves working on a mathematics problem, as well as discussing students' thinking about mathematics. Writing is used regularly as a tool for pushing the group's thinking forward. While the sessions are not scripted, they are – in many

important ways -- well-laid out “textbooks” for professional development leaders.

Consider a typical problem teachers might mull over:

Delving Into Student Thinking

Teacher Liz Sweeney asked her fifth grade class to come up with methods for solving some two-digit multiplication problems. Jemea solved 29×12 correctly when she rounded the single factor, 29, to 30. She added twelve 30s to get 360 and then subtracted 12 to get 348.

Thomas used a rounding strategy also, but his led him astray. Setting out to simplify the problem, 36×17 , he added 4 to the 36 to get 40 and 3 to the 17 to get 20. That left him with a multiplication problem he could solve easily: $40 \times 20 = 800$. Thomas then subtracted the 4 and the 3 from 800 to arrive at his answer: 793. (The correct answer is 612.)

Ms. Sweeney found Thomas's strategy intriguing and asked him to present it to the class. The students saw right away that Thomas had reached an incorrect answer, but they couldn't understand why. The class spent the next two sessions investigating multidigit multiplication as they searched for Thomas's error.

They came to see that when Thomas added 4 to 36 and 3 to 17, he changed the problem to $(36 + 4) \times (17 + 3)$, which equals: $36 \times 17 + 4 \times 17 + 3 \times 36 + 4 \times 3$. Thomas had added in a lot more than the 4 and the 3 he later subtracted.

In between sessions, participating teachers try out ideas and strategies in their own classrooms, complete homework assignments, and bring those experiences back to the professional development seminars to further the group's work. Homework

assignments include reviewing curricular materials and analyzing the central mathematical concepts, reading and re-reading case materials, and posing questions to the teachers' own students.

There are also supports and materials designed for facilitators. One important support is the facilitator's guide. This guide describes each session in detail and the facilitator is given instructions about how to prepare, as well as what materials to bring. Each session is laid out in considerable detail. Each guide also includes narrative descriptions of each session taken from "Maxine's Journal." Maxine is a fictional facilitator, created from a composite of experiences that facilitators reported during the piloting of each module. The journal includes descriptions of Maxine's goals, her frustrations and "aha" moments, where people sat, what they said, how teachers struggled with the mathematics. Also included are examples of how Maxine responded to teachers' writing and homework.

Another support offered by the developers is an intensive two-week summer DMI Leadership Institute. Although such training is not required, many facilitators have attended these institutes. Training is module-specific, that is, facilitators work through the modules that they will teach. Beginning facilitators are taught by seasoned ones. The emphasis is on both experiencing the DMI materials as a learner and then analyzing the experience of learning as a facilitator. Through this process they learn how to use the materials, typical problems they might encounter, and alternative strategies leaders might use to stimulate teacher learning. They discuss how to handle common issues that arise in professional development, especially professional development that challenges teachers to improve their knowledge of mathematics (which can be emotionally and

intellectually trying). After training, facilitators return to their home school districts and collaborate with others in offering the module(s) that they have been trained to teach.

As already noted, at the local level, DMI is often adapted to the needs and contexts of a particular school district or collaborative. These local adaptations can involve changing schedules for how and when DMI is offered, combining these professional development efforts with others offered in the school district, aligning mathematics professional development with curriculum reform, and the like. The degree to which these adaptations alter teachers' opportunities to engage DMI concepts and materials, we may expect the program to be more or less effective at promoting teacher learning.

Background²

In light of current educational reforms that endorse and promote high-stakes assessment and accountability, teachers are under growing pressure to change and improve their practice (Borko, 2004; Cohen & Ball, 1990; Sykes, 1996). With the intense focus on raising student achievement, teacher quality is seen as critical. For student achievement to improve, classroom instruction must improve, and thus the responsibility for change falls squarely on the shoulders of teachers. It has also become increasingly clear that the requisite changes are substantial in nature (Cohen & Ball, 1990). Instead of simply adopting a new technique here or implementing a new program there, teachers are being asked to make deep changes in their instructional practice. To meet the ambitious content standards of current educational reforms, teachers must examine -- and often transform -- their core beliefs, knowledge, and habits (Gallucci,

2003; Stein, Smith, & Silver, 1999). This commitment to fundamental teacher change is very much at the heart of DMI.

For teachers to make these substantial changes in their practice, they must have access to sufficient opportunities to learn. There is a general sense among researchers and educators that teachers' opportunities to learn are shaped by program features and by the facilitators leading those programs (e.g., Borko, 2004; Cohen & Hill, 2000; Franke & Kazemi, 2001; Kennedy, 1999; Stein, Silver, & Smith, 1998; Stein, et al., 1999; Wilson & Berne, 1999).

However, the empirical evidence to buttress these core assertions is still not sufficiently extensive or rigorous. Researchers and scholars have nominated characteristics of effective programs that seem to provide such opportunities to learn (e.g., Borko, 2004; Franke & Kazemi, 2001; Kennedy, 1999; Wilson & Berne, 1999). Out of this research have come numerous sets of core principles of effective professional development (see Elmore, 2002; Hawley & Valli, 1999; Stein, et al., 1999; Stigler, & Hiebert, 1999). Although each framework is slightly different, they tend to share common features. Weiss and Pasley (2006) summarize these features:

According to this consensus view, high-quality PD programs are grounded in research and clinical knowledge of teaching and learning. They are aligned with a school's curriculum and assessments and focused on student learning in that setting. They facilitate teachers' collaboration both within and across schools, they use existing teacher expertise to plan activities and cultivate leaders, and they include mechanisms for garnering principal support. High-quality PD programs both model and explicitly discuss methods of good practice (such as inquiry-

based methods in science) and provide teachers with active learning opportunities. These programs aim to build teachers' content knowledge and pedagogical skills. Finally, they are intensive, sustained over time to allow for integration of new knowledge into practice, and include follow-up support. (pp. 1-2)

Thus, DMI's principles, content, and structures nicely align with these characteristics. DMI is content-specific and informed by the latest research. Everything is organized around students' mathematical work. Teachers examine curricular materials and assessments, and the seminars involve collaborations within and across schools (and sometimes districts). Facilitators are taught to model instructional methods that teachers might use (when appropriate) with their students, teachers are provided with opportunities to work through mathematical problems designed to engage teachers as learners, discuss students' work with colleagues, investigate student thinking in their own classrooms, and explore mathematics, student learning, and educational research through reading and writing assignments. The work is aimed at increasing their knowledge of both mathematics and how students think about specific mathematical concepts. Finally, the sessions are intense, and scheduled to take place over time with multiple opportunities for follow-up and the integration of classroom experience with seminar discussions.

Returning to the literature, less research has focused on the characteristics or preparation of effective facilitators. However, case study evidence suggests facilitators matter in important ways (e.g., Stein, et al., 1998; Stein, et al., 1999). For example, effective facilitators must be able to clearly articulate the goals of the professional development program to participants. Facilitators must be able to recognize teachers' learning challenges and know how to productively guide teachers toward deeper

understandings. Facilitators must also be able to adapt the program to meet local needs, but know how to avoid adaptations likely to decrease the efficacy of the program.

Finally, facilitators must be able to develop communities of practice in which it is safe and expected for teachers to be learners. These tasks are not straightforward and dramatically shift the traditional role of “professional developer.” We note again that although we use “must” here, this argument remains largely theoretical. Although it seems logical that facilitators matter to teacher learning, we have very little evidence about the degree to which facilitators shape teacher learning (Borko, 2004).

Here too DMI’s materials for facilitators appear well aligned with current best practice. Facilitators’ materials are well-specified and built on the structures necessary for increasing the fidelity of the implementation. Facilitators are provided with both written materials and training seminars in learning mathematics, facilitating discussions, listening for teachers’ concerns and confusions, and providing helpful and targeted feedback. As teachers themselves most facilitators do not have much experience teaching other teachers. Moving beyond opportunities to casually discuss students and mathematics to focused, intensive discussions meant to change teachers’ understandings of those things requires considerable knowledge and experience. Facilitator experience matters. The DMI developers have engaged this challenge directly, and thus the facilitator materials and training suggest that DMI is a particularly good site to investigate the relationships among teacher learning, facilitator experience, and program features in a program consistent with the consensus view of effective professional development.

Mathematical Knowledge for Teaching

While DMI's goals extend beyond increasing teachers' mathematical knowledge, this inquiry focused on that aspect of the work. Many policymakers and reformers are interested in professional development that is "content-rich" and there has been a great deal of concern about the content preparation of teachers generally (Allen, 2003; Paige, 2002). In the case of mathematics, part of the concern is rooted in research, which has documented the lack of mathematical knowledge in prospective elementary teachers (Ball, 1990; Ball, Lubienski, & Mewburn, 2001; Ma, 1999). Some concern, however, is also rooted in a general perception that subject matter knowledge *should* matter (no matter what the empirical evidence says) and that many teachers are not serious scholars of the content they teach. Considering how central this concern is, it is worrisome that there is so little solid data to inform these discussions. In the most recent review of research on teachers' undergraduate disciplinary majors, Floden and Menketti (2005) found that there is very little research on teachers' subject-specific study except in mathematics. The research in mathematics suggests that teachers' study of mathematics shows a positive correlation with high school pupils' mathematics learning (Allen, 2003; Floden & Menketti, 2005; Wilson, Floden, & Ferrini-Mundy, 2001).

While there is limited empirical evidence, there has been considerable conceptual work in this domain. Shulman and his colleagues (1986, 1987; Grossman, 1990; Wilson, Shulman, & Richert, 1987) initially proposed that teachers needed both subject matter knowledge and pedagogical content knowledge (PCK). The former was similar to the content learned in university disciplinary departments, the latter entailed knowledge of how to represent the content for students, as well as knowledge of how students understand, misunderstand, or typically interact with school knowledge. Ma (1999) dug

further into the concept of mathematical knowledge, arguing that the difference between U.S. and Chinese teachers was that the latter possessed a “profound understanding of fundamental mathematics”, thereby articulating some of the characteristics of teachers’ content knowledge. Ball, Hill, Bass, and their colleagues (Ball & Bass, 2000, 2003; Ball, Hill, & Bass, 2005; Hill, 2004) have taken the argument even further, hypothesizing that there is a mathematical knowledge for teaching (MKT) that is distinct from mathematical knowledge that one might acquire while majoring in mathematics or learning mathematics for other forms of professional work, like engineering³. This argument has been picked up and developed by others in the mathematics and mathematics education community (e.g., Wu, 2005, 2006, 2007). Supporting this conceptual work, MKT has recently been demonstrated to be related to student achievement and mathematics instruction (Blunk, 2007; Hill, Blunk, et al., 2007; Hill, Rowan, & Ball, 2005).

Research on and the Measurement of Teacher Learning

In sum, there is considerable agreement about what professional development ought to look like, and equally considerable work on the kinds of mathematical knowledge teachers need. Yet research that examines whether this kind of professional development leads to increased teacher or student learning is still needed (Borko, 2004; Desimone, Porter, Garet, Yoon, & Birman, 2002; Desimone, Porter, Birman, Garet, & Yoon, 2002; Garet, Porter, Desimone, Birman, & Yoon, 2001; Hill & Ball, 2004; Wilson & Berne, 1999; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007).

One important complication in remedying this problem is that the technology available to measure teacher learning is limited. Traditionally, professional development programs have asked teachers to participate in surveys, or exit polls about the relative

merits of the professional development. Some professional development programs have attempted to use locally developed pre- and post-tests of teacher outcomes. When one looks to teacher education programs for measures used to assess what graduates know, there is also very little available by way of common, validated measures of what teachers learn. Researchers have used a variety of different proxies, many of which are relatively gross indicators: college majors, grade point averages, retention in teaching (Wilson, et al., 2001). Teacher tests are equally problematic (Wilson & Youngs, 2005). Thus, for professional development programs interested in collecting professionally responsible, publicly credible evidence of what teachers learn in a particular professional development program, there is very little by way of trustworthy methods or measures at their disposal.

Researchers working on the Study of Instructional Improvement (SII) have done groundbreaking work in this regard (e.g., Hill, Ball, Blunk, Goffney, & Rowan, 2007; Hill, Schilling, & Ball, 2004; Rowan & Ball, 2004), building on earlier work in assessing teachers' mathematical knowledge (e.g., Kennedy, Ball, & McDiarmid, 1993; Rowan, Chiang, & Miller, 1997). Because some aspects of teacher knowledge appear to be domain-specific, these assessments require the generation of hundreds of items that are then used to create scales.⁴ Part of that development work has entailed sharing those measures with other projects, thereby gathering more data on teachers' responses, which – in turn – leads to more refinement of the measures, and improved data on reliability and validity. The research we report on here took advantage of this opportunity and used SII items. The use of these items provided us one way to assess the external validity of any changes in teachers' knowledge we might see on DMI-specific measures.

In sum, in the past five years, the calls have become more insistent for research on professional development and its effects (e.g., Borko, 2004; Wilson & Berne, 1999; Yoon, et al., 2007). While such research is beginning to amass, the field still needs empirical studies of professional development. In particular, there is a need for much more detailed information on the factors that shape the effects and effectiveness of professional development programs. DMI is a professional development program designed to align with many of the ideals expressed in the 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 that may be able to support the type of intensive, sustained, collaborative inquiry that is seen as crucial to effective professional development. Thus, it is a professional development program ideally positioned as a research site to investigate the relationships between teacher learning, facilitator experiences, and the program features.

Method

The research reported here is based upon a NSF-funded evaluation of DMI that was designed to provide summative information and analysis on DMI's effects. The study involved ten DMI sites. The research was designed to compare the knowledge of teachers who participated in DMI with comparable teachers who did not. The design was nested, involving two levels with documentation at the site and teacher level. We do not attempt to build a complete model of teacher learning using hierarchical linear modeling (HLM), but we do 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 two).

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 getting DMI started. We explained the details of the research design (see below), and invited sites to participate. Each site had a local study facilitator who administered instruments. We eventually established relationships with ten 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 case of South Hadley, MA, there are two separate cohorts of teachers taught by two different sets of facilitators. Within the ten sites there were two types: sites in which teachers were all from the same district and sites in which teachers came from different districts to a regional professional development center. Both South Hadley⁵ sites, Hudson, and Bryant were regional professional development centers.

DMI at the Sites

There was little variability across sites in terms of how DMI was structured and carried out. Most sites (9 of 11) scheduled meetings after school every week or every other week for between 7 and 16 weeks. This was true for both BST and MMO. One site used a summer format which lasted one week and met intensively each day. And one site used a mixture of these two formats. Teachers were unpaid volunteers in all but four of the sites. No teachers were required to take BST or MMO. All facilitators reported that they assigned the prescribed homework to teachers and most sites (9 of 11) reported that they assigned homework for every session of BST and MMO. The same proportion

of site facilitators also gave written feedback on those homework assignments on a weekly basis. The regular assignment of homework and written feedback from facilitators is viewed by DMI developers as crucial to participants' learning.

As described earlier, the consensus view of effective professional development posits that teachers should be in robust communities of practice that engage them around issues relevant to their teaching practice. All of the 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 which 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 teaching practices of their teachers. From facilitators' perspectives, these sessions went well, produced learning, and met participants' needs.

Facilitators' survey responses suggest that DMI was implemented with a high degree of fidelity. There were small variations in implementation but core aspects of the seminar remained intact. These core aspects of the seminar are the features of effective professional development, and as such, confirm that these sites are reasonable locations to investigate the relationships between teacher learning and facilitator experiences.

Participants

All teachers were voluntary participants in DMI or peers of these teachers who volunteered to serve in the "non-DMI" comparison group. The non-DMI group of teachers taught at the same site but was not matched by grade level, amount of teaching experience or the like. Matching participants in that way is difficult and resource-intensive. We settled for a comparison group of elementary teachers who taught in the

same school district(s) or region. The DMI and non-DMI groups of teachers are described in more detail below.

Three hundred and eleven teachers participated in the research. The majority were female (DMI 90%; non-DMI, 87%), and Caucasian (DMI 81%; non-DMI 85%). DMI participants' average years of teaching experience was 12.6 ($SD = 8.75$), and non-DMI 12.92 ($SD = 10.28$). The majority of participants were elementary school teachers, although a few participants were administrators/specialists, special education, or middle or high school teachers.

[Table 1 about here]

As Table 1 details, DMI participants had a greater proportion of administrators and secondary teachers. 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, 46% non-DMI). Prior to the study, both DMI and non-DMI participants reported similar numbers of math-specific professional development hours, with the exception of a small cohort of DMI teachers who had more than 81 hours of math pd in the past three years.

Data on the schools in which participants taught was not consistently reported by participants, but it appears that both DMI and non-DMI teachers taught in schools with varying levels 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.

Despite the differences noted above, the DMI and non-DMI groups were similar to one another in terms of their experience, teaching contexts, and the roles they played in those contexts.

These comparison groups and the use of pre-tests of MKT allow us to eliminate two potential alternative hypotheses that might explain any changes in knowledge we might measure. One alternative hypothesis suggests any changes in teachers' knowledge might be the result of other district initiatives, not DMI. Using the comparison group, we can be confident that concurrent initiatives -- differentiated instruction, data-driven decision making, or the professional development that accompanies textbook adoption -- were not influencing teacher learning.⁶ We can presume that because the teachers are in the same district, they experience those initiatives in similar ways and therefore, the main difference between the groups is the decision to participate in DMI or not.

We should note however, that some site facilitators indicated it was harder to recruit non-DMI teachers to the project. All but two facilitators indicated it was easy or moderately easy to recruit non-DMI teachers. In the two more difficult sites, facilitators felt non-DMI teachers were "tired" of professional development or simply had little interest in math professional development specifically. Though it seems reasonable to presume the non-DMI group was less enthusiastic about math professional development (otherwise they might have volunteered for DMI), as will be detailed later in the paper, there was no difference between the two groups' MKT at the beginning of the study.

A second alternative hypothesis suggests that unobserved differences that cause some teachers to volunteer for DMI and other teachers not to volunteer may account for any learning gains we might see between the two groups of teachers. Teaching

experience, number of previous math professional development hours, and other similar characteristics might cause teachers who elect to participate in DMI to learn more than participants without those characteristics. We cannot rule this out entirely; however, by using the pre-test measures, we can be reasonably confident that while motivation, effort, interest in mathematics may differ, the groups did not begin with different levels of MKT.

Pre- and Post-Tests of Teacher Knowledge⁷

At each site, all participating teachers in DMI took pre- and post-tests designed to assess their mathematical knowledge for teaching and their pedagogical content knowledge. The teachers took the pre-test prior to participating in one module, Building a System of Tens (BST), and then the post-test after completing a second, Making Meaning for Operations (MMO). We measured teachers' learning after both DMI courses 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 the national standards.

Items for these tests were developed in two ways: multiple-choice items were selected from the item bank developed by SII staff. Open-ended items were developed

from assessments used by DMI project leaders in early stages of the development of these seminars and the associated materials. We describe each briefly.

Multiple-Choice Items. We created an instrument using items developed by SII (Hill, et al., 2004). Several criteria were used to select appropriate items. First, we selected items that were within the general domain of mathematical topics covered by DMI. Second, we selected items that were the most reliable (based on previous IRT analyses), as well as items that represented a range of difficulty levels (based on those same analyses). We then eliminated some questions that would lead to overrepresentation of some content, and focused on questions that required analysis of student work, instructional tasks, representation tools, and content. When we believed that we had a defensible, representative set of items, the DMI advisory board -- composed of mathematicians, DMI developers, professional development leaders, and mathematics educators and education researchers -- reviewed the items for content, relevance, and face validity.⁸

Open-Ended Items. We also developed a set of open-ended items. The items were based on an existing instrument originally created by the DMI authors. That original instrument evolved over ten years as an embedded assessment used in some DMI sessions. The items were examined against the relevant BST and MMO learning goals in the domains of pedagogical content knowledge (although some also appear related to mathematical knowledge for teaching). Additional items were added to improve coverage of content represented in the learning goals for the DMI modules.

Items were also reviewed by the DMI advisory board. Through this process, items were changed, modified, or completely replaced. Our records show that we went

through 12 distinct 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.

The scoring rubric went through a lengthy development process. Near final versions were reviewed for content validity by the DMI authors and the Professional Development Study Group (a group of researchers and practitioners from EDC, TERC, TUFTS, Mt. Holyoke, and the greater Boston area that meet regularly to share work in progress).

In the end, our multiple choice assessment contained 20 questions, which are scored dichotomously. Therefore, scores on this assessment can range from a low of 0 to a high of 20. The internal consistency reliability estimates for the pre and post multiple choice assessments were .79 for the pretest and .80 for the posttest. The open-ended assessment contains 14 items that are scored with a rubric, and scores can range from 0 to 32. The internal consistency reliability estimate for the open-ended pretest was .76. At post-test, the internal consistency reliability estimate was .79. Table 4 contains information regarding the subscale reliabilities and average inter-item correlations.⁹

Approximately 11% of the open-ended responses (n=34) were scored by two different raters to assess the degree of interrater agreement. The mean inter-rater agreement across the 14 open-ended items was 80.8%. These estimates of score reliability indicate that the pre- and post- measures of mathematics knowledge exhibited acceptable internal consistency, stability, and consistency across raters. For the purposes of the HLM analyses, we used the first thirteen questions on the multiple choice pretest, since most of the teachers from site 1 took a version of the multiple choice pretest that

did not contain questions 14 and 15. However, the multiple-choice posttest used for the HLM analyses contained all 15 questions.

[Table 2 about here]

In addition, we conducted a confirmatory factor analysis of the assessment scores using MPLUS. We estimated a two-factor model, in which open-ended responses were predicted by an open-ended factor and multiple choice responses were predicted by a multiple choice factor. While the two-factor model exhibited reasonable fit for both the pre-test data ($\chi^2=197.45$ with 167 degrees of freedom, $p=.054$), the correlation between the two latent factors was almost perfect ($r=.96$). However, given our interest in potential differences in the two types of assessments, we decided to maintain two separate scales for our HLM analyses: a multiple choice scale and an open-ended scale.

Surveying Cross Site Variability

While we were interested in the effects of DMI across all sites, we were also 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©. 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. We worked with the study administrator to identify the relevant personnel. One study administrator, nine study administrator/facilitators, six facilitators, and four district

personnel completed the questionnaire. Two of the 22 individuals asked to participate did not respond.

For this analysis, we elected to focus on variation across facilitators as our level-2 variable. We presumed that teacher learning would be enhanced if facilitators were more skilled and knowledgeable about DMI. This assumption, as pointed out earlier, resonates both with current best practice in professional development more generally and with DMI in particular (recall the showcased voice of Maxine and her journal in the supporting materials). 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, and so we went in search of a responsible proxy.

Based on our interviews with site personnel, it was clear that facilitators regarded as “excellent” by their peers had many different kinds of experiences working with the DMI materials. It seems likely that if a facilitator has had many different experiences with DMI, she – like Maxine -- is more likely to have seen the range of ways teachers engage the material. More experienced facilitators are also more likely to have worked with the materials longer and therefore have higher levels of relevant knowledge.¹⁰ 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 questionnaires, facilitators were asked to respond to several questions regarding their opportunities to learn DMI. By summing nine questions that assessed the

breadth of DMI experiences facilitators reported we constructed the OTL variable. These questions asked about whether the facilitator had attended the leadership institute at Mt. Holyoke College, attended a leadership institute at another location, apprenticed to another facilitator, co-facilitated, taught at the Mt. Holyoke leadership institute, participated in a study group of other facilitators, 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 co-facilitated, most (8 of 11) 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 of the facilitators went to the Mt. Holyoke institute, had been in a facilitator study group, or participated in another leadership training program. Just three of the facilitators had taught at Mt. Holyoke or written cases about their teaching experiences.

Data Analysis

To evaluate the efficacy of the DMI training, we compared the post-test scores of teachers who received DMI training to scores of comparison teachers who were not involved in the training, after controlling for pre-test scores. To analyze the effectiveness of the intervention, we conducted a series of multilevel regression analyses using HLM 6.03. We analyzed the impact of the professional development training on two separate dependent variables: multiple choice test score and open-ended test score. For the purposes of the analyses, level 1 was the teacher level; level 2 was the site level. We estimated all models using restricted maximum likelihood estimation, given the small level-2 sample size ($N=10$). For the analyses of each of the dependent variables, we used

a model building approach (Raudenbush & Bryk, 2002). First, we estimated the unconditional model. Then we estimated a level-1 model, which included two independent variables: treatment (coded 0 for comparison and 1 for treatment), pretest score, which was grand mean centered, and the interaction between pretest and treatment. At first, we allowed the slopes of the intercept and the two level-1 slopes to vary. Next, we eliminated any level-2 random slope effects that were not statistically significant (we always allowed the intercept to randomly vary across site), and we compared this simpler, less parameterized model to the full model, using a chi-square difference test. If the chi-square test suggested that the simpler model with fewer random effects provided no worse fit to the data than the more parameterized model, then the simpler model was retained. Finally, we estimated a level-2 model, with OTL 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

Sample

The final sample consisted of 311 teachers from 10 sites. Nine of the 10 sites (n=259) contained both treatment and comparison teachers. The tenth site (n=51) was a training of trainers site, which included only treatment teachers. We began data gathering at 11 sites. After receiving pre test data from one of these sites, the site dropped out of the study. For six 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 post test compared to 55% from the comparison group). The remaining two sites had high attrition rates. At one of the these sites, the attrition rate was 46% and due to record keeping problems it is unclear whether the losses were equally distributed between the DMI and comparison groups or not. The final site also had problematic attrition rates, but due to ID assignment problems (this was the first site to begin testing) it is unclear exactly how many participants were lost between the pre and post sessions due to attrition as opposed to IDs that could not be matched.

Thus, there was great variability in the final sample sizes across the 10 sites. Whereas, the smallest sites (sites 4 and 10) consisted of only 14 teachers, the largest site (site 6) contained 66 teachers. Table 3 shows the sample sizes and the means and standard deviations for the treatment and comparison groups by site for the pre and post assessments.

[Table 3 about here]

To ensure that there were no differences between the two groups at pretest, we compared the means of the two groups on the two pre-assessments. We excluded the trainer site from these analyses, as we expected that participants would be higher at pretest than the other sites, and this site contained only treatment teachers.¹¹ There were no statistically significant differences between the treatment group and the comparison group on either of the two assessments given at pretest. It is impossible to know whether the two groups of teachers were equivalent prior to the start of the training. However, similarity of the two groups on their pretest scores suggests that the two groups had

similar skill levels on the material covered in the DMI training. The demographic information reported in the method section further supports the claim of group similarity.

Multiple Choice Assessment

After controlling for pretest scores on the multiple-choice assessment, there were statistically significant differences between the two groups in terms of their posttest scores on the multiple choice assessment (Table 4). After controlling for pretest scores, the DMI group outperformed the comparison group by .99 points. The Cohen's *d* effect size for the magnitude of this difference (adjusting for pretest scores) was .26 standard deviation units, a small effect. Facilitators' OTL were not a statistically significant predictor of the treatment slope. Although the interaction between the pretest and the treatment was not statically significant, the *p*-value approached statistical significance ($p=.08$). Therefore, we chose to leave this term in the model for completeness. The coefficient for the interaction was negative (-.14), indicating that the relationship between pretest scores and post-test scores was stronger for teachers in the comparison group than it was for teachers in the treatment group. After controlling for the other variables in the model, for every point higher that a comparison teacher scored at pretest, his or her predicted score at post test increased by .88 points. In contrast, for every point higher that a treatment teacher scored at pretest, his or her predicted post-test score increased by .74 points. After controlling for treatment, pretest, and the interaction between treatment and pretest, no statistically significant between-site variability in multiple choice post-test scores remained to be explained.

[Table 4 about here]

Open-Ended Choice Assessment

After controlling for pretest scores on the open-ended choice assessment, there were statistically significant differences between the two groups in terms of their posttest scores on the open-ended assessment (Table 5). After controlling for pretest scores, the DMI group outperformed the comparison group by 3.15 points. The Cohen's d effect size for the magnitude of this difference (adjusting for pretest scores) was .49 standard deviation units, a moderate effect size. Additionally, the site facilitator's OTL was a statistically significant predictor of the treatment slope. After controlling for the other variables in the model, for every one point increase in facilitator's breadth of knowledge, there was corresponding .42 unit increase in the predicted posttest score for a treatment teacher. In other words, the expected difference between the treatment and comparison groups was larger in sites where facilitators had a broader range of opportunities to learn about DMI. Although the interaction between the pretest and the treatment was not statistically significant, the p -value approached statistical significance. Therefore, we chose to leave this term in the model for completeness. The coefficient for the interaction was negative (-.17), indicating that the relationship between pretest scores and post-test scores was stronger for teachers in the comparison group than it was for teachers in the treatment group. After controlling for the other variables in the model, for every point higher that a comparison teacher scored at pretest, his or her predicted score at post test increased by .83 points. In contrast, for every point higher that a treatment teacher scored at pretest, his or her predicted post-test score increased by .66 points. After controlling for treatment, pretest, and the interaction between treatment and pretest, no statistically significant between-site variability in open-ended post-test scores remained to be explained.

[Table 5 about here]

Comparison of Multiple Choice and Open-Ended Assessments

While the DMI group did exhibit higher scores than the comparison group on both of the 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 were in the same mathematical vicinity as the DMI content; the open-ended items were designed to assess DMI specific learning goals in the domain of pedagogical content knowledge. In addition, while facilitator's breadth of DMI 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.

Discussion

Through participation in high fidelity implementations of DMI, teachers at ten sites across the country deepened their mathematical knowledge. This finding was consistent on two assessments--one generalized instrument that reasonably matches the content in the DMI modules and one instrument more tailored to the articulated learning goals for the particular modules under study, but limited in interpretation because its validity has not been tested outside the context of the current inquiry. The degree to which teachers developed their mathematical knowledge was mediated by their facilitators' learning experiences. This finding supports theory-based arguments and case study evidence regarding the importance of facilitators in the efficacy of professional development programs.

The claims we can make about what the improvement in teachers' knowledge might mean for teaching practices and student achievement 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 SII items have considerable validity evidence associated with them. In their three original forms, the items have been found to be related to student achievement (Hill, et al., 2005) and to teaching practices (Blunk, 2007; Hill, Blunk, et al., 2007). We used a modified form of the items in order to assure alignment 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 of the items may be related to changes in other measures of teaching and learning.¹² Future work should examine these relationships more directly. Our current work only looks at one potential arena of impact. There may (or may not) be impacts in other areas such as teaching practices and/or student achievement.

The use of both the open-ended and multiple choice measures was deliberate. We wanted to have some evidence that changes we saw on the open-ended items were related to other knowledge measures that have strong external and internal validity data. The finding that DMI teachers learned more on both measures provides some convergent validity evidence that the open-ended items are measuring changes in knowledge that are related to the changes SII has been able to detect. This gives us confidence the changes we see on the open-ended items are real and worthy of further investigation.

Ours is necessarily a modest study. Nonetheless, we have some conjectures about why DMI participants outperformed their colleagues. These results may have arisen because a small cadre of the participants (23%) participated in extensive professional

development prior to the start of the study (see Table 1). Though there were no differences between DMI and non-DMI participants' mathematical knowledge prior to the study, it is possible that unobservables associated with prior math professional development might make that small cadre of participants particularly able to learn from DMI. While this is certainly a concern, the lack of difference on pre-tests, and the small proportion of participants who had extensive experiences, gives us confidence the effects we see are robust.

An alternative explanation suggests that the larger percentage of secondary teachers (3%) and administrators, specialists or non-classroom teachers (12%) in the DMI group, as compared to the non-DMI group, might positively bias the results because modules on numbers and operations could be viewed as a somewhat artificial treatment. Anecdotal evidence from the sites suggests secondary mathematics teachers were grateful for the opportunity to engage such content because they regularly must address students' fundamental mathematical understandings (such as the ones covered in BST and MMO) if they are to successfully teach "secondary" content. If the DMI treatment were artificial in some way, we would not expect secondary teachers to engage the content so eagerly. It is possible that administrators, specialists, and non-classroom based teachers have more extensive mathematical knowledge. We would expect significant knowledge differences between these subgroups and the rest of the teachers to appear on pre-test measures. No such differences were found.

A final alternative explanation suggests that given the similarity in the demonstrated knowledge of DMI and non-DMI participants prior to the study, there may be something about the qualities and structure of DMI that made a difference. DMI has

features consistent with the features other researchers have found to be associated with effective professional development. DMI is well specified at the teacher and facilitator level. It is a coherent, long term, and narrowly focused approach on particular topics in K-8 mathematics. DMI 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. Our evidence (at least in the form of facilitator self-report) suggests these features were implemented with high degrees of fidelity. 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 have the opportunity to learn more in practice than anywhere else. 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 an enormous 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 mathematical understanding and pedagogical content knowledge into their classrooms, and try things out. Repeatedly teachers told us of the revelations (and we use that language intentionally here, for teachers often used almost evangelical language to describe their learning (Mattson, 2003)) they experienced – 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 S. Cohen's (2004) case study research.

Our findings also suggest that the broader a facilitator's opportunities to learn about DMI, the more teacher learning she was able to facilitate. We hypothesize this effectiveness with teachers likely 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 or 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. This kind of knowledge might have enabled facilitators' ability to adapt the modules to their context and participants.

Of course, the study design and size does not allow us to test these conjectures, and so we will have to leave it for future research to mount the large-scale, resource-rich efforts necessary to test such hypotheses. Given the lack of good instrumentation for assessing teacher or facilitator knowledge, and a lack of variation in how DMI was implemented at the sites, we can only hypothesize about the relationships among facilitator experiences, teacher learning, and program features.

Implications

As previously discussed, one goal of this work was to understand what it takes to move from locally designed and administered assessments of professional development

to more robust, defensible measures that connect to existing understandings of teachers' mathematical knowledge for teaching. A second goal was to document the relationships among teacher learning, facilitators, and program features. This paper has made progress on both goals, but it is worth emphasizing some of the challenges researchers face as they set about doing this work.

These kinds of studies are understandably rare. They require collaboration, both between professional development leaders and researchers, but also among the various sites with one professional development "system." We had the support of DMI developers who provided invaluable assistance and access to sites. This support made the study possible. Collaborations are necessary to the conduct of professional development studies but they require individuals with different needs, agendas, and audiences to work successfully together over lengthy periods of time. Professional associations and graduate schools should pay additional attention to understanding these issues and developing researchers' skills in these areas.

These types of studies are expensive--financially and politically. Locating sites, negotiating access, collecting the data, and creating a database are resource-intensive tasks, fraught with potential pitfalls. We paid participants, site administrators, and scorers. 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. While 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 good will in order to sustain data collection.

Further, studies like this are designed in the research world where one can lay out and control all the study's demands. But when working with schools and districts reality intervenes. Site administrators change jobs or leave the district. Sites get restructured and undergo political turmoil. Communication fails. Gathering valid, reliable data within the given budget and time frame requires constant attention and problem solving. The judgment required for this problem solving is often invisible in formal reports of research and therefore, can go unexamined. Given the importance of this judgment to the successful completion of this type of work, we should consciously broaden these conversations beyond the relatively small group of senior researchers conducting much of this work to include graduate students as well as funders and practitioners.

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 ten years ago. In those intervening 10 years, developers got smarter about the requirements of effective DMI training, incorporating those lessons in subsequent drafts of modules. They developed long standing relationships with educators around the county. 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 ten years of work made this study possible. This is a point that was recently 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 funders 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 the calls for the documentation of teacher learning in professional development settings will not be answered in the short term. It also suggests 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. We found that DMI participants demonstrated significantly increased knowledge as assessed by both the multiple choice assessment and the open-ended assessment, although the strength of that relationship was stronger for the 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 as well. The learning we document 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.

This effort moves beyond analyses of a single successful professional development program in one site to explore the elements and resources that contribute to effective professional development programs at scale. There is still however, much work to be done on both DMI and other professional development programs. As 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 contributes in important, if limited, ways to our need to have better empirical evidence for grounding claims about high quality professional development.

References

- Allen, M. (2003). *Eight questions on teacher preparation: What does the research say?* Denver, CO: Education Commission of the States.
- American Federation of Teachers. (2002). *Principles for professional development: AFT's guidelines for creating professional development programs that make a difference.* Washington, D.C.: Author.
- Author (2007).
- Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *Elementary School Journal*, 90, 449-466.
- Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on the teaching and learning of mathematics* (pp. 83-104). Westport, CT: Ablex.
- Ball, D. L., & Bass, H. (2003). Toward a practice based theory of mathematical knowledge for teaching. In B. Davis & E. Simmt (Eds.), *Proceeding of the 2002 Annual Meeting of the Canadian Mathematics Education Study Group* (pp. 3-14). Edmonton, AB: CMESG/GCEDM.
- 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. & Bass, H. (2005). Knowing mathematics for teaching. *American Educator*, 29(3), 14-22, 43-46.

- Ball, D. L., Lubienski, S. T., & Mewborn, D. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433-456). New York: Macmillan.
- Blunk, M. L. (2007). *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.
- Cohen, D. K., & Ball, D. L. (1990). Policy and practice: An overview. *Educational Evaluation and Policy Analysis*, 12, 233-239.
- Cohen, D. K., & Hill, H. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record*, 102(2), 294-343.
- Cohen, S. (2004). *Teachers' professional development and the elementary mathematics classroom: Bridging understanding to light*. Lawrence Erlbaum, Mahwah: NJ.
- The College Board. (2006). *Teachers and the uncertain American future*. New York: Author.
- Davenport, L. & Morse, A. (2001). *Fostering a stance of inquiry among teachers: Professional development in mathematics education* (Paper #13, Center for the Development of Teaching's Paper Series). Newton, MA: EDC.
- Desimone, L., Porter, A. C., Garet, M., Yoon, K. S., & Birman, B. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, 24, 81-112.

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.

Elmore, R. F. (2002). *Bridging the gap between standards and achievement: The imperative for professional development in education*. Washington, D.C.: Albert Shanker Institute.

Franke, E., & Kazemi, E. (2001). Teaching as learning within a community of practice. In T. Wood, B.S. Nelson, and J. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics*. Mahwah, NJ: Lawrence Erlbaum Associates.

Floden, R. E., & Menketti, M. (2005). Research on the effects of coursework in the arts and sciences and in the foundations of 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: Lawrence Erlbaum Associates.

Gallucci, C. (2003). Communities of practice and the mediation of teachers' responses to standards-based reform. *Education Policy Analysis Archives, 11*(35). Retrieved 10 January 2006 from <http://epaa.asu.edu/epaa/>.

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.

- Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education*. New York: Teachers College Press.
- Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 127-150). San Francisco: Jossey-Bass.
- Hill, H. C. (2004). Professional development standards and practices in elementary school mathematics. *Elementary School Journal*, 104, 215-231.
- 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 Perspectives*, 5(2-3), p.107-118.
- Hill, H.C., Ball, D.L., Sleep, L. & Lewis, J.M. (2007) Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts? In F. Lester (Ed.), *Handbook for research on mathematics education* (2nd ed), p. 111-155. Charlotte, NC: Information Age Publishing.
- Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., et al. (2007). *Mathematical knowledge for teaching and the mathematical quality of instructions: An exploratory study*. Unpublished manuscript.

- Hill, H., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Education Research Journal*, 42, 371-406.
- Hill, H., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, 105, 11-30.
- Kennedy, M. (1999). *Form and substance in in-service mathematics and science programs*. National Institute for Science Education, University of Wisconsin.
- 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*. Hillsdale, NJ: Lawrence Erlbaum.
- Mattson, S. (2003). *A changing metaphor: Instructional reform as evangelism*. Unpublished dissertation, Michigan State University.
- National Research Council. (2003). *Strategic education research partnership*. Washington, D.C.: Author.
- Paige, R. (2002). *Meeting the highly qualified teachers challenge: The Secretary's annual report on teacher quality*. Washington, DC: U.S. Department of Education.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models (2nd ed.)*. Newbury Park, CA: Sage

- Remillard, J. T., & Geist, P. (2002). Supporting teachers' professional learning through navigating openings in the curriculum. *Journal of Mathematics Teacher Education, 5*(1).
- Rowan, B., & Ball, D. L. (2004). Introduction: Measuring instruction. *Elementary School Journal, 105*, 3-10.
- Rowan, B., Chiang, F., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on student achievement. *Sociology of Education, 70*, 256-284.
- Schifter, D., Bastable, V., Russell, S. J., with Cohen, S., Lester, J. B., & Yaffee, L. (1999). *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). *Making meaning for operations, casebook*. Parsippany, NJ: Dale Seymour.
- 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), 1-22.
- Stein, M. K., Silver, E. A., & Smith, M. S. (1998). Mathematics reform and teacher development: A community of practice perspective. In J. Greeno & S. 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.
- Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: The Free Press.
- Sykes, G. (1996). Reform of and as professional development, *Phi Delta Kappan, 77*, 464-467.
- Weiss, I. R., & Pasley, J. D. (2006). *Scaling up instructional improvement through teacher professional development: Insights from the local systemic change initiative*. Research report of the Consortium for Policy Research in Education (R8-44). Philadelphia: CPRE.
- 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., Duffy, H., Fiori, N., Halladay, J., & Mapuranga, R. (2006). *Assembling a good team: Teacher learning from professional development*. Final report to the Noyce Foundation. East Lansing, MI: Michigan State University.
- Wilson, S. M., Floden, R., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations*. Seattle, WA: University of Washington Center for 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.
- 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: Lawrence Erlbaum Associates.
- Wu, H. (2005). *Key mathematical ideas in grades 5-8*. Paper presented at the annual meeting of the National Council of Teachers of Mathematics. Anaheim, CA.
- Wu, H. (2006). *Professional development: The hard work of learning mathematics*. Presentation at the fall southeastern section meeting of the American Mathematical Society, Johnson City, TN.
- Wu, H. (2007). *Mathematics for K-12 teaching*. Unpublished manuscript.
- Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & 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 from <http://ies.ed.gov/ncee/edlabs>.

Table 1

Participants' educational roles, previous math 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 non-classroom role	13	1
Middle or high school teacher	11	8
Hours of math 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 FRL eligibility was 49 and 44% for DMI and non-DMI groups, respectively.

Table 2

Internal Consistency Reliability (Cronbach's Alpha)

Test	# Items	Mean inter-item correlation (IIC)	Standard Deviation of IIC	Cronbach's Alpha
Multiple Choice Pre-test	18	.15	.095	.77
Multiple Choice Post-test	20	.15	.089	.79
Open-ended Pre-test	14	.19	.126	.76
Open-ended Post-test	14	.222	.122	.79

Table 3

Descriptive Statistics for the 10 DMI Sites

Site			MC	MC	OE	OE
			Pretest	Post	Pretest	Post
1.00	Comparison	Mean	10.83	12.92	13.83	14.50
	n=12	Std. Deviation	2.55	3.78	4.95	4.42
	DMI	Mean	10.24	12.18	16.41	17.12
	n=17	Std. Deviation	3.47	3.07	4.90	6.12
2.00	Comparison	Mean	8.50	10.40	12.50	14.70
	N=10	Std. Deviation	4.14	4.48	4.14	4.47
	DMI	Mean	9.58	10.79	12.79	15.84
	N=19	Std. Deviation	4.27	4.76	5.66	7.40
3.00	Comparison	Mean	8.00	9.75	11.44	10.44
	N=16	Std. Deviation	3.81	3.71	4.18	3.63
	DMI	Mean	9.71	12.71	14.06	15.53
	N=17	Std. Deviation	4.34	3.53	9.01	7.40
4.00	Comparison	Mean	11.40	13.60	16.50	16.44
	N=10	Std. Deviation	2.55	2.63	5.04	7.57
	DMI	Mean	10.75	14.50	15.00	16.75
	N=4	Std. Deviation	1.26	1.00	2.94	5.91
5.00	Comparison	Mean	8.75	9.85	13.35	13.80
	N=20	Std. Deviation	3.96	4.26	5.12	6.49
	DMI	Mean	10.38	12.81	13.90	19.62
	N=21	Std. Deviation	2.69	3.30	5.52	5.31
6.00	Comparison	Mean	12.00	14.78	19.11	18.22
	N=9	Std. Deviation	3.04	2.17	4.20	6.16
	DMI	Mean	12.00	14.33	16.33	20.50
	N=6	Std. Deviation	3.22	3.39	3.88	5.96
8.00	Comparison	Mean	8.06	9.05	11.14	11.78
	N=36	Std. Deviation	3.45	3.46	5.22	5.86
	DMI	Mean	9.10	11.90	11.00	16.00
	N=30	Std. Deviation	3.33	3.65	4.71	5.64
9.00	Comparison	Mean	10.88	11.50	15.00	15.13
	N=8	Std. Deviation	3.45	4.21	6.76	7.49

Measuring the Effects of Professional Development 50

	DMI	Mean	9.00	12.13	13.50	15.25
	N=8	Std. Deviation	3.21	3.04	6.05	6.86
10.0	Comparison	Mean	10.13	10.50	11.38	14.50
	N=8	Std. Deviation	2.80	2.73	2.67	2.78
	DMI	Mean	10.00	12.57	9.29	16.29
	N=6	Std. Deviation	3.52	2.23	5.96	3.25
11.0	DMI	Mean	12.84	14.40	18.17	21.79
	N=51	Std. Deviation	2.39	3.13	5.54	5.89
total	Comparison	Mean	9.29	10.72	13.08	13.56
	N=129	Std. Deviation	3.65	3.96	5.28	5.89
	DMI	Mean	10.69	12.88	14.69	18.27
	N=179	Std. Deviation	3.48	3.56	6.26	6.51

Table 4

Final HLM for Post Multiple Choice Score

Fixed Effects	Coefficient (SE)	<i>t</i> (df)	<i>p</i>
Model for intercept (β_0)			
Intercept (γ_{00})	11.47 (.22)	51.90 (9)	<.001
Model for TRT slopes (β_1)			
Intercept (γ_{10})	.99 (0.29)	3.45 (305)	.001
Model for MC PRETEST slopes (β_2)			
Intercept (γ_{20})	.88 (.06)	14.91 (305)	<.001
Model for PRE X TRT slopes (β_3)			
Intercept (γ_{30})	-.14 (.08)	-1.76 (305)	.08
Random Effects (Variance Components)			
Between Site Var. in intercepts (τ_{00})	0.006	χ^2 (df) 11.25 (9)	<i>P</i> .26
Var. within sites (σ^2)	5.90		

Table 5

Final HLM for Post Open-ended Scores

Fixed Effects	Coefficient (SE)	<i>t</i> (df)	<i>p</i>
Model for intercept (β_0)			
Intercept (γ_{00})	14.33 (.47)	30.34 (9)	<.001
Model for TRT slopes (β_1)			
Intercept (γ_{10})	3.15 (.54)	5.84 (306)	<.001
OTL (γ_{11})	.42 (.17)	2.52 (306)	.013
Model for PRETEST slopes (β_2)			
Intercept (γ_{20})	.83 (.08)	11.08 (306)	<.001
Model for PRE X TRT slopes (β_3)			
Intercept (γ_{30})	-.17 (.09)	-1.86 (306)	.06
Random Effects (Variance Components)			
Between site var. in intercepts (τ_{00})	.57	χ^2 (df) 14.90 (9)	.09
Var. within sites (σ^2)	19.37		

Appendix

Correlations Among Multiple-Choice and Open-Ended Subscales at Pre-Test and Post-Test

	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).

Author Note

This paper reflects a team effort. Authors contributed equally to the research and/or writing. This work was supported by a grant from the National Science Foundation under Grant No. ESI-0242609. Any opinions, findings, conclusions, or recommendations expressed in this paper 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 for Instructional Improvement for their wise counsel concerning the assessment of teacher knowledge.

Notes

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.

² For more detail on this literature, see Wilson, Duffy, Fiori, Halladay, & Mapuranga (2006).

³ See Hill, Ball, Sleep, and Lewis (in press) for a comprehensive recounting of the evolution of this line of work.

⁴ Heather Hill (personal communication) estimates the development costs for the modest number of items that currently exist to be between five and seven million dollars, which includes seven years investment in both piloting the items and their validation.

⁵ One of the South Hadley sites was a “training of trainers” site —participants who were learning DMI in order to teach others. This group was only added after a site we had gathered pre test data from was unable to continue with the study.

⁶ These and other initiatives occurred during the study period at district sites.

⁷ For a more detailed explanation of the measures development, see Author (2007).

⁸ This process is explained in greater detail in Author (2007).

⁹ The correlation between the pretest administration of the multiple-choice assessment and the post-test administration of the multiple-choice assessment was .77. The correlation between the pretest administration of the open-ended assessment and the post-test administration of the open-ended assessment was .67 (see Appendix). While dissatisfying as a result, this is as one would expect given the fuzzy 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).

¹⁰ 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 with what facilitators need.

¹¹ Because of the uniqueness of this site, we could have dropped it from all subsequent analyses. We elected not to do so after running analyses with and without the site and finding consistent results.

¹² Confirmatory factor analyses (documented in Author (2007)) suggest that at a minimum, our form of the items has some amount of internal validity.