Professional Development Research: Consensus, Crossroads, and Challenges

Heather C. Hill, Mary Beisiegel and Robin Jacob

EDUCATIONAL RESEARCHER 2013 42: 476 originally published online 7 November 2013
DOI: 10.3102/0013189X13512674

Professional Development Research: Consensus, Crossroads, and Challenges

Heather C. Hill[1], Mary Beisiegel[2], and Robin Jacob[3]

Commentaries regarding appropriate methods for researching professional development have been a frequent topic in recent issues of *Educational Researcher* as well as other venues. In this article, the authors extend this discussion by observing that randomized trials of specific professional development programs have not enhanced our knowledge of effective program characteristics, leaving practitioners without guidance with regard to best practices. In response, the authors propose that scholars should execute more rigorous comparisons of professional development designs *at the initial stages* of program development and use information derived from these studies to build a professional knowledge base. The authors illustrate with examples of both a proposed study and reviews of evidence on key questions in the literature.

Keywords: educational policy; experimental research; professional development; program evaluation

The field of professional development research has reached a crossroad. Through studies conducted over the past two decades, scholars have identified program design elements thought to maximize teacher learning, including a strong content focus, inquiry-oriented learning approaches, collaborative participation, and coherence with school curricula and policies (e.g., Cohen & Hill, 2001; Garet, Porter, Desimone, Birman, & Yoon, 2001; Penuel, Fishman, Yamaguchi, & Gallagher, 2007). Agreement about this list had reached a level such that many in the field felt comfortable characterizing support for the list as a "consensus" (Desimone, 2009; Penuel et al., 2007; Russell, Kleiman, Carey, & Douglas, 2009). Yet disappointing results from recent rigorous studies of programs containing some or all of these features have turned this consensus on its head (Arens et al., 2012; Bos et al., 2012; Garet et al., 2008; Garet et al., 2011; Santagata, Kersting, Givven, & Stigler, 2011). At the same time, recent econometric studies of professional development have generally indicated weak return from district dollars invested in this sector (Harris & Sass, 2011; Jacob & Lefgren, 2004; for an exception, see Angrist & Lavy, 2001).

It is too early to tell why these results—and especially the results of randomized trials—contradict conventional wisdom among researchers. The content of the specific programs evaluated may have been ineffectual, or programs may have deviated from best practices in important ways due to poor implementation or difficulties scaling the program to multiple sites. Poor research design—inadequate measures, insufficient power—may also contribute to these findings. However, it is not too early—and it is in fact critically important at this crossroad—to re-evaluate the research paradigm in professional development. Going program by program and—often at great expense—conducting large-scale evaluations involving multiple measures of teaching and learning has not, to date, resulted in an accretion of credible, usable knowledge within the professional development and practitioner community. Yet developers and policymakers urgently need more rigorous evidence that describes how professional development design elements impact the likelihood of program success (Dede, Ketelhut, Whitehouse, Breit, & McCloskey, 2009). This is particularly important as most professional development is home-grown; it arises from district or local developers' needs and interests, has a relatively short shelf-life, and proceeds with little or no formal evaluation.

This article suggests a new approach to research on professional development. This approach is based on the idea that scholars should execute more rigorous comparisons of professional development design elements *at the initial stages* of program development. The designs compared must be carefully linked to open questions within the professional development literature, allowing the field to effectively accumulate evidence on issues of importance to local providers. This initial work must also progress with multiple groups of teachers and multiple facilitators, lest idiosyncratic results at one location lead

[1]Harvard Graduate School of Education, Cambridge, MA
[2]Oregon State University, Corvallis, OR
[3]University of Michigan, Ann Arbor, MI

Educational Researcher, Vol. 42 No. 9, pp. 476–487
DOI: 10.3102/0013189X13512674

developers to incorrect conclusions about program design and promise. Finally, as studies accumulate, analysts should conduct meta-analyses that inform these open questions.

In what follows, we provide background on the history of professional development research and describe the trends toward large-scale experimentation that have occurred in recent years. We then propose a lower cost yet rigorous alternative and demonstrate how this might work by presenting examples of two important stages of this work.

Four Decades of Research on Professional Development

Research on professional development has changed considerably over the past few decades. Prior to 1990, evaluators generally investigated the effectiveness of professional development on a small scale, often using teacher reports of change or satisfaction as a primary criterion of program success (Frechtling, Sharp, Carey, & Vaden-Kiernan, 1995). Although this likely remains the dominant method for the evaluation of many local programs, it has been largely supplanted in the mainstream research literature due to three factors.

The first is the development of more objective measures with which researchers can gauge professional development effects. In some part, these measures have been a byproduct of increased testing in U.S. public schools; with yearly test data, as is required by No Child Left Behind, researchers can estimate whether a professional development program results in higher teacher "value added" scores, or the classroom-average improvement in test scores between adjacent years' tests adjusted for student and peer characteristics (e.g., Harris & Sass, 2011; Jacob & Lefgren, 2004). In another part, federal and foundation funding priorities have led to the creation of more direct measures of teacher knowledge and classroom practice (e.g., Borko, Stecher, Alonzo, Moncure, & McClam, 2005; Bush, Ronau, Brown, & Myers, 2006; Carlisle, Kelcey, Rowan, & Phelps, 2011; Grossman et al., 2010; Learning Mathematics for Teaching Project, 2011; McCrory, Floden, Ferrini-Mundy, Reckase, & Senk, 2012; Smith & Banilower, 2006). These measures have become widely used in program evaluations of professional development (e.g., Heller, Daehler, Won, Shinohara, & Miratrix, 2012).

Second, some professional development scholars have sought to compare the effects of program *features*, rather than evaluating specific programs. These researchers rely primarily on surveys that ask teachers to report on both the content of their professional development experiences as well as key outcomes, such as their knowledge, perceived teaching capacity, or instructional practices. To identify best practices in professional development, researchers then compare variability in program content with variability in these outcomes (see, e.g., Cohen & Hill, 2001; Garet et al., 2001; Penuel et al., 2007; for a research paradigm built upon this strategy, see Desimone, 2009). Results from this literature tend to point in the same direction, toward the use of novel professional development structures (e.g., teacher study groups, coaching), collaboration among colleagues within schools, a subject matter focus rather than a focus on generic teaching practices, and fostering teacher learning by engaging in active tasks, such as curriculum design, enactment, and reflection. However, research in this tradition can rarely make strong causal statements about these features due to at least two design flaws. First, selection effects—teachers intentionally choosing programs that match their preexisting instruction or disposition to change—may lead to correlations between professional development characteristics and better instructional or student outcomes absent an actual causal impact. Second, this literature relies on teacher self-report rather than objective measures of instructional or student-level outcomes.

Partially in response, a third strand of scholarship in this field has focused on research designs that feature random assignment of teachers or schools to treatment condition. Under the right conditions, random assignment allows scholars to make causal inferences regarding program effects; these inferences are not possible in nonexperimental designs (Feuer, Towne, & Shavelson, 2002). The earliest of these studies were carried out in the late 1970s, with results demonstrating that programs promoting highly organized, direct instruction techniques positively impacted student outcomes (Good, Grouws, & Ebmeier, 1983). In the 1980s and 1990s, researchers followed with several random-assignment evaluations of content-specific programs, such as Carpenter, Fennema, Peterson, Chiang, and Loef's (1989) *Cognitively Guided Instruction*. Results from these studies foreshadowed many of the best practices identified in the survey-research tradition, although conclusions from them required a higher level of inference: Because studies evaluated single programs that were combinations of many discrete elements, it was difficult to discern which among those elements—or which interactions among elements—led to program success.

As research methods for studying professional development changed, two other important developments occurred in the field. First, Hilda Borko published a highly influential article delineating a three-phase approach to studying professional development (Borko, 2004). In this approach:

> Phase 1 research activities focus on an individual professional development program at a single site. Researchers typically study the professional development program, teachers as learners, and the relationships between these two elements of the system. The facilitator and context remain unstudied. In Phase 2, researchers study a single professional development program enacted by more than one facilitator at more than one site, exploring the relationships among facilitators, the professional development program, and teachers as learners. In Phase 3, the research focus broadens to comparing multiple professional development programs, each enacted at multiple sites. Researchers study the relationships among all four elements of a professional development system: facilitator, professional development program, teachers as learners, and context. (Borko, 2004, p. 4)

Borko's idea of studying a program first at a single site and then in multiple locations has been frequently adopted in practice (e.g., the progression of research described in Daehler & Shinohara, 2001; Heller et al., 2012; Shinohara, Daehler, & Heller, 2004).

A second important development occurred during the George W. Bush administration, when the U.S. Department of Education established the Institute for Education Sciences (IES) and reoriented education research priorities away from developmental,

descriptive, and survey-based research and toward inquiries built around random assignment studies (U.S. Department of Education, 2002). To do so, IES adopted a goal structure for competitive grants. Although this goal structure has changed slightly over the years, its major milestones remain the same:

- *Goal 1: Exploratory.* Generating hypotheses or theories from existing datasets (e.g., exploring Early Childhood Longitudinal Study data).
- *Goal 2: Development and innovation.* Developing interventions that demonstrate a positive effect on student outcomes, often in a single pilot study setting.
- *Goal 3: Efficacy and replication.* Implementing successful Goal 2 interventions in authentic yet favorable settings (e.g., with extra support from developers, sites where teachers are interested in the intervention) to determine their effects on student outcomes, usually across multiple sites.
- *Goal 4: Effectiveness.* Determining whether successful Goal 3 interventions continue their success under conditions of routine practice (e.g., without special assistance by the developer) and at multiple sites (IES, 2012).

In Goal 2 onward, evaluating the intervention vis-à-vis student outcomes is a requirement; IES has also indicated that it prefers randomized or strong quasi-experimental designs, even in Goal 2. Contract research within IES's National Center for Educational Evaluation (NCEE) and awards to the Regional Educational Laboratories (RELs) were also redesigned to feature causal research, with at least eight cluster randomized trials of professional development launched from these agencies over the years 2005–2011. In recent years, the National Science Foundation (NSF) has adopted a similar, although less rigid, structure for both its Discovery Research K–12 (DRK–12) and Research and Evaluation on Education in Science and Engineering (REESE) programs.

An examination of results from the first several years of studies funded under the new IES goal structure suggests effects vary widely. Although some IES-funded cluster randomized trials find effects of professional development on student outcomes (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Landry, Anthony, Swank, & Monseque-Bailey, 2009; Penuel, Gallagher, & Moorthy, 2011; Powell & Diamond, 2011; Ramey, Ramey, Crowell, Grace, & Timraz, in press), for many others, findings are either null or largely null (Buysse, Castro, Peisner-Feinberg, 2010; Cabalo, Ma, & Jaciw, 2007; Gersten, Dimino, Jayanthi, Kim, & Santoro, 2010; Santagata, Kersting et al., 2011). Furthermore, an examination of studies contracted by the NCEE (Garet et al., 2008; Garet et al., 2011) and the RELs finds more mixed results. In the REL case, only two of six studies yielded positive effects on student outcomes (Finklestein, Hansen, Huang, Hirschman, & Huang, 2011; Newman et al., 2012), and one of those studies (Newman et al., 2012) showed an effect size of only .05 standard deviations or a difference of about two percentile points in favor of treatment group teachers. In the NCEE case, studies of four different mathematics and reading programs showed no effect of professional development on student outcomes (Garet et al., 2008; Garet et al., 2011). NCEE and REL studies are seldom conducted by the developers of the program and often use distal indicators of outcomes, suggesting these studies may show typical program effects under routine conditions, such as limited support from developers and typical levels of district participation and support.

These results are also notable in the sense that many of the programs studied contained key elements thought to enhance professional development outcomes, such as generous time in professional development, a content focus, active learning opportunities, collaboration among teachers, and collective participation within schools (Garet et al., 2001). There are many reasons why these studies may have failed to find impacts, including ineffective content, poor or incomplete program implementation, inadequate statistical power, poor measures (e.g., lack of alignment between intervention and outcome measures), problematic randomization, or improper data analysis. However, it seems likely that, in at least in some cases, poorly designed programs also contributed to these disappointing findings. At an average cost of several million dollars per study, disappointing results in so many studies suggest that future dollars might be better spent exploring which design features lead to promising professional development outcomes, rather than exclusively evaluating already-established programs.

The Crossroad

Against this backdrop, we argue that it is time to reevaluate recommendations for conducting research in the field of professional development.

One reason lies in a commonsense limitation to single-program research studies: that such studies assess a *package* of professional development, not its specific features (Wayne, Yoon, Zhu, Cronen, & Garet, 2008). If positive outcomes occur, it is difficult to determine what specific features—or combination of features—led to program success. If positive findings are not found, it is also difficult to identify why. Although meta-analyses of multiple programs with varying characteristics are possible, the field may take years to develop a large enough sample of studies for proper analysis. For example, Kennedy (1999) analyzed only 12 programs and Yoon, Duncan, Lee, Scarloss, and Shapley (2007) identified only nine studies that met more stringent criteria for inclusion in such analyses.

A second reason to reassess the current paradigm relates to the fact that most professional development is locally developed and implemented and that the life cycle of any particular approach or program is relatively short. One reason for this is that professional development is often seen as the chief vehicle for implementing new policy initiatives—data-driven instruction, the Common Core State Standards, individualized instruction, integrating technology into instruction, or motivational/aspirational programs. With a shifting policy agenda, professional development must be frequently designed and redesigned to meet teachers' and districts' needs. In this situation, guidance for developers regarding best design practices—rather than a list of programs "that work"—becomes critical.

Third, developers of professional development at any level, be it in a university or in a school district office, have questions that now go beyond established conventional wisdom. These questions have arisen in part because technology affords new

Table 1
Professional Development Research Stages

Stage	Description	Indicators	Comparable Current IES Goal
Stage 1	One-site pilot	Teacher and developer perceptions	—
Stage 2	Randomized controlled trial holding content the same but varying features of program delivery	Proximal, low-cost measures of teacher knowledge and practice Teacher and developer perceptions	—
Stage 3	Efficacy trial of moderate size	Standard measures of teacher knowledge and practice Student outcomes	Goal 3
Stage 4	Scale-up trial	Standard measures of teacher knowledge and practice Student outcomes	Goal 4
Stage 5	Meta-analysis of Stages 2, 3, and 4 studies	—	—

professional development formats and practices. For instance, in recent years, the wide availability of videotape technology has led to interest in video clubs, particularly for the study of mathematics instruction (see, e.g., Santagata & Angelici, 2010; Sherin & van Es, 2009). Prospective designers of such professional development face multiple decisions for which there is little solid evidence. Is it more effective for teachers to view their own videotape or stock footage from other teachers' classrooms? Is it better for teachers to develop their own method for analyzing the videotape or to use an externally imposed lens? How much should a facilitator intervene during discussions to move teachers toward a specific view of a video? We argue that the answers to each of these questions will both profoundly shape the program's model and also condition its success. Doubtlessly, designers of coaching, online experiences, and data-study programs have similar questions.

A fourth reason to reevaluate the cycle of research in professional development lies in the length of time needed to collect information on whether even one program is successful. The typical length of an IES grant is between 3 and 4 years, suggesting that a progression from Goal 2 (*Development*) to Goal 4 (*Effectiveness*) would take roughly a decade to complete. This, and the fact that it is only across multiple such programs of research that conclusions can be drawn about effective program features, suggests that rigorously derived answers to questions about effective program features may yet be far off.

Finally, in the IES model, research starts with a small-scale demonstration project, often at a hand-selected site. This means both that estimates of the likelihood of program success and the baking of "promising" features into the program design are decided based on the reaction of a relatively small group of teachers to content conveyed by a small team of facilitators. However, positive effects may also result from characteristics of the original group of teachers, characteristics of the facilitators, or the interaction of the two. These facilitators are often the developers of the program, which may additionally positively bias results. Thus, given perennial issues with "scaling up" interventions within the U.S. educational system (Elmore, 1996), it may be unwise to wait until the program is in final form to cross sites and employ multiple facilitators; developing knowledge about how the program varies across sites and facilitators early in the design process may improve the robustness of the program.

We argue that this critique holds several lessons for the design of professional development studies going forward. Designers must early on test their programs in multiple contexts and with multiple facilitators. Such studies should be as rigorous as possible, and their results should help answer broad questions regarding the design of professional development, especially those that are currently relevant to local program designers. We continue below by sketching out a new paradigm for this kind of research.

The Challenge

Our proposal centers on the idea that the field of professional development research can execute more rigorous, cross-site research at early stages of program development.

Through such programs of research, which we illustrate in more detail below, we believe we can generate usable knowledge for the field and improve the likelihood that professional development will positively impact instruction and student outcomes.

Like both Borko (2004) and the IES goal structure, we argue that professional development should proceed in several specific stages (see Table 1).

During Stage 1, we propose first a brief one-site pilot to ensure the feasibility of the program—in other words, will the intended intervention work with real teachers, or is it unrealistic in its expectations? During the pilot, changes in program features could be assessed in successive sessions or with subgroups of teachers, with new permutations and adaptations emerging via feedback from both teachers and developer observation. Importantly, developers may wish to test-drive the program features they will investigate further in Stage 2. Developers may want to work in a school or district with ideal conditions—supportive administration, common planning time, alignment of curriculum, and assessment with professional development. This one-site pilot need not take much time—perhaps four to six sessions—and in many cases will not require grant funding, as data collection is solely informal feedback and the sample of teachers would naturally be quite small. This stage could be

undertaken within a single academic year for costs that range in the tens of thousands of dollars, most of it professional developer and teacher time, some of which could be donated or paid for via existing professional development funds. For instance, a developer may conduct several coaching cycles with a select group of teachers, comparing unstructured coaching to coaching based on an observational instrument, and receiving feedback from those teachers about which has the greatest likelihood of becoming fruitfully integrated into district routines and, ultimately, affecting practice and learning.

Stage 2 would involve a randomized clinical trial that holds the basic program *content* constant, varies the *features* of delivery, and then searches across multiple sites for impacts on logically important but proximal program outcomes. Three steps characterize research at this stage. We describe each, demonstrating through a hypothetical example how they might work.

The first step would be identification of critical program design questions, either through reading the research literature or analyzing design possibilities. For instance, developers may be interested in whether feedback to teachers based on an observational instrument such as the Danielson *Framework for Teaching* (*FFT*) can be effective in improving teaching and, ultimately, learning. However, developers may have questions as to whether this feedback must be delivered via individual, in-person coaching; whether it can be delivered to grade-level teams in a group setting; or whether it can be delivered to individuals using videotape and remote coaching. The latter two would present cost savings, particularly in large or rural districts. A review of the literature would confirm that there has already been interest in the remote versus face-to-face option, and based on this and cost considerations, the developers would identify these three features to test.

Notably, we argue that for this strand of research, developers keep the content of the professional development—the materials, resources, and activities that form the basis for teacher learning—fixed. In line with our review of current professional development research, we expect content may consist of intended instructional practices, protocols for analyzing data from student assessments, or how to deploy a new set of curriculum materials. Although this content could be permuted in a Stage 2–like setting, doing so would shift the question to the efficacy of the *content*, not the *features*, of the professional development. Because we advocate for building generalizable principles for effective professional development design, we suggest that permutations of content be dealt with separately, in another line of research.

The second step in Stage 2 would be the provision of professional development with each feature to multiple groups of individuals by multiple facilitators. This breaks the dependence between outcomes and the specifics of any group, facilitator, or group–facilitator combination. To continue the example above, the professional development provider would recruit a group of teachers from multiple schools or even districts to participate in the study. These teachers would then be randomly assigned to either a control (no treatment) or one of the three *FFT* treatment conditions. We recommend randomly assigning teachers rather than schools or teaching teams in order to maximize the power of the study to detect effects. Teachers could be blocked by school, grade level, or teaching team to further maximize the power of the study.[1] Teachers within each random assignment group would then be randomly divided into smaller groups (e.g., 5 groups of size 15) for the provision of the professional development, and the three versions of the program, identical save for delivery method, would be provided intensively over the course a short time span, perhaps 6 months. Figure 1 provides a visual overview of this design.

Finally in this second stage, researchers would use proximal, low-cost outcomes to gauge the initial success of the program. Because classroom practice is still expensive to capture and measure at scale, Stage 2 researchers should identify a logic model that specifies the relationship between the program, mediators, and outcomes—for example, program content leads to changes in teacher knowledge, skills, and habits of mind, which leads to changes in instruction and ultimately student outcomes—and then measure the most proximal indicators of learning from the program. Importantly, these mediators cannot purely consist of self-reports. In the above example, for instance, researchers may wish to gauge teachers' ability to critically analyze and reflect upon classroom instruction, and researchers may also want to assess whether teachers deepen their understanding of the observational instrument itself. Programs that intend to improve teachers' mathematical knowledge for teaching (Bush et al., 2006; Hill, Schilling, & Ball, 2004) or efficacy (Tschannen-Moran & Hoy, 2001) might use measures of these constructs, as such measures have been linked to student outcomes. Participation patterns—gauged through analyses of teachers' talk—can also become an important outcome, in that investigators can seek to understand which features lead to meaningful discussion and teacher contributions (see, e.g., Sowder, Philipp, Armstrong, & Schappelle, 1998). In all cases, we would recommend modeling change—that is, collecting data at baseline, midpoint, and the conclusion of the professional development, then assessing whether individuals have changed as a result of their progression through the program. An important criterion is the degree of alignment between program intent and these outcomes; poor alignment would generate false negative outcomes and hinder attempts to learn across studies. All this information could be collected either online or during the professional development sessions themselves.

Under some conditions, it may also be possible to use student achievement data on state assessments[2] to assess the impact of the various versions of the professional development program on achievement (see Jacob, Goddard, & Kim, n.d., for a full discussion of the use of aggregate data in evaluating school or grade-level interventions). However, interpreting such results would depend upon design and power considerations (discussed below). In instances where using state assessments to measure potential impacts is not feasible or possible, researchers could examine the correlations between student outcomes and the proximal outcomes the intervention is attempting to change. If those proximal outcomes are in fact predictive of student outcomes across teachers in the study, this would suggest tentative support for researchers' causal model. Although such analyses would be exploratory, results might suggest moving forward with an efficacy trial and could potentially identify the most promising design features—that is, the ones that produce the

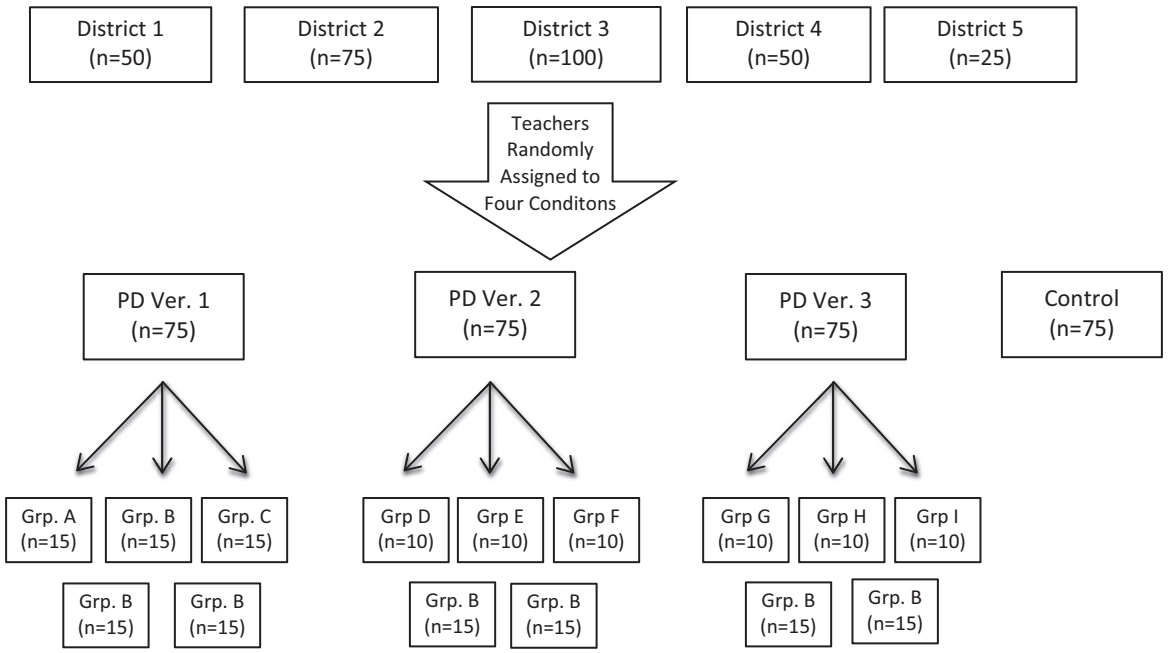

FIGURE 1. *One potential Stage 2 design*

largest impacts on the proximal outcomes most highly correlated with student outcomes.

During Stage 2, teachers' own voices would again be important in helping developers determine best-case designs for their intervention. Although teacher opinions about professional development are not generally thought to provide adequate evidence for a summative evaluation, in this case teachers would be providing feedback about how specific program elements either encouraged or discouraged the intended outcomes, about logistical difficulties involved with mounting the program in schools, and about the likelihood of success on a wider scale.

As noted above, we argue for random assignment of teachers to treatment condition at this stage. Random assignment is difficult to achieve under the best of circumstances; with a large number of sites and at such an early stage of development of the program, some researchers may be satisfied with establishing baseline equivalence among groups. However, we argue that when possible, random assignment should be sought, as it makes the analysis and interpretation of results far more clear; in many situations, the added effort for random assignment will likely be small relative to these gains in analyses and interpretation.

At the end of Stage 2, the developer would test for effects in the treatment conditions vis-à-vis the control condition, inspect for differences in effects across treatment conditions (e.g., "Does in-person delivery provide better average outcomes than online delivery?"), integrate the more descriptive data from teachers into these assessments, weigh the cost of each design, and incorporate any available data on student achievement outcomes. Using this information, developers can make decisions about the design that maximizes the likelihood of effecting change in teaching and learning and that minimizes the costs of implementing the program. For example, if online delivery is equally effective, but less costly than face-to-face delivery, program developers may choose to move forward with an online-only program. Similarly, if investigators do not observe large differences in teacher outcomes between the professional development variations, the least costly version could be implemented, and all three treatment groups could be combined and compared to the control group, with substantially more power, to assess the overall impact of the program on the outcomes of interest. If no differences are found between any of the treatment groups and the control group, then program developers would need to assess whether it was the delivery of the professional development (PD) or other aspects of the program that were ineffective.

In many cases, the data collection costs for Stage 2 studies would be quite low—primarily the costs associated with developing and implementing an online tool for collecting teacher data and providing teachers with a small monetary incentive (e.g., $40) for completing the survey or data collection instrument. We estimate that for a 6-month program, the individual-random-assignment version of this plan could be done for as low as $300,000, including the cost of subsidizing teacher release time, depending on the exact sample size and details of the data collection. Given the very high costs of many current randomized control trials, this would be a cost-efficient way to obtain valuable information.

In Stage 3, developers would modify and finalize the intervention materials or protocols and then take them into a traditional efficacy trial. It would be important at this stage to conduct a random assignment study comparing control and treatment groups and to examine the efficacy of the program vis-à-vis both standard teaching and learning outcomes, such as observations of practice and intervention-sensitive assessments. As Borko (2004) and IES both recommend, this would also be a time to widen the pool of facilitators beyond the group that conducted the Stage 2 piloting. However, this work would be

conducted at a limited number of sites (schools or districts) in order to facilitate the collection of rich data on mediators and outcomes.

If successful, a program would enter into Stage 4, where scale-up trials would track the effects of the program as disseminated under standard (market) conditions. Stages 3 and 4 are comparable to IES Goals 3 and 4 in length and cost; experience suggests that such studies can be completed in 2 to 3 years, depending on the desired length of the professional development program and how prepared materials and facilitators are for each trial. These studies typically cost between $1 million and $3 million apiece, depending again on the length of the program and the extent to which mediators (e.g., instruction) and alternative outcomes (e.g., nonstandardized tests) are measured.

Stage 5 of our approach would involve collecting evidence from Stages 2, 3, and 4 trials and conducting either meta-analyses or structured reviews of these findings. For instance, a scholar interested in whether professional development delivered online is as or more effective than in-person delivery would identify and collect the reports on the *FFT* study suggested above. This scholar would also collect information from other, similar studies, ultimately synthesizing these findings as well as analyzing other program-related information (size, facilitator expertise) that would guide the field in future professional development design.

During Stage 5, scholars could also investigate the effect of contextual factors on professional development effectiveness. For instance, the individual conducting a review of in-person versus online professional development may group studies by content, finding that the online condition works well for some teaching skills but not others. Alternatively, a feature like group feedback on videos of instruction may work in schools with well-functioning professional communities but not in schools without such communities; taking care to measure such key attributes at the program outset could contribute to advancing the field in this way.

For this stage of the research approach to be viable, the field would have to agree on important features to permute in such studies. However, we suspect that within the field, several practical questions already stand out, including the question of in-person versus online delivery of content, the appropriate role of the facilitator, and methods for connecting program content to teachers' everyday practices, for instance via coaching. Several commentaries and reviews of research (Desimone, 2009; Hawley & Valli, 1999; Kennedy, 1999) have also suggested theoretical features that bear investigation. The field may also have questions specific to particular modes of professional development, such as content learning or video study. Conducting multiple studies of these issues satisfies many scholars' recommendations that research hypotheses be subject to repeated tests and that results should generalize and replicate across studies (National Research Council, 2002). Once completed, these structured reviews on specific research questions could be available to professional development practitioners intending to design new programs. Such an approach could easily fit within a small grant for meta-analysis, as has historically been supported by NSF's REESE and DRK-12 programs, and could be completed in 1 to 2 years of working time.

Such an approach to professional development research would have several advantages. First, developers would be able to rigorously test various adaptations to the program during the development phase, meaning that the ultimate program entered into a Stage 3 study would be one that promises maximum impact on classroom instruction and student learning. We suspect that many programs now in existence would be effective, or more effective, had the developer been able to carefully examine the effects of alternative designs on outcomes. Second, even in Stage 2, each condition would be implemented in different sites by multiple facilitators. This ensures that program effects are not site- or facilitator-specific; it would also have the effect of providing information about the ability of a program to scale across sites. Third, the use of proximal outcome variables simplifies, to a degree, the kinds of analyses that would need to be conducted, making this kind of study accessible to those without expertise in value-added modeling or similar methods. Fourth, because of its limited data collection plans, Stage 2 research is reasonably cost-efficient and would fit well within a modest-sized grant proposal. Finally, because of the inclusion of a traditional control group, there would be evidence of program effects over the set of proximal outcomes, effects that would suggest whether or not to take the intervention into a Stage 3 study.

Stage 2 Sample Size and Logistical Considerations

Above, we argue that the costs for Stage 2 studies would be smaller in comparison to a full-scale randomized control trial examining student outcomes. In part, this is due to reduced data collection burden. In another part, this is because in many cases, such studies could be embedded in existing professional development experiences or evaluations. Finally, because such studies are aiming to detect large changes in proximal outcomes, they can be powered to detect larger rather than smaller effect sizes.

In fact, we recommend that researchers power their studies to detect effects in the range of .30 standard deviations. Without relatively large impacts on mediators, one is unlikely to see an impact on more distal outcomes such as student achievement. Previous studies have demonstrated impacts between treatment and control groups on proximal outcomes such as teacher knowledge or specific instructional practice in the range of .30 to .50 standard deviations without demonstrating impacts on student achievement (e.g., Garet et al., 2008; Garet et al., 2011). Furthermore, researchers have also demonstrated impacts based on variations in treatment delivery (e.g., professional development with or without coaching) of over .20 standard deviations without impacting student achievement (Garet et al., 2008). Thus, searching for impacts on proximal outcomes that are less than .30 standard deviations, we believe, is of limited value.

Detecting differences in effects between conditions will require a considerable investment in both researcher and teacher time. In a scenario in which individual teachers are randomly assigned to various design permutations, we estimate that approximately 75 teachers per random assignment group would be needed to detect effects of around .30 standard deviations. This assumes a two-tailed test with an alpha of .10,[3] a pretest predictor correlated with the outcome at around .70 and power equal to .80. With four

groups (three treatments and a comparison group), the total sample size would be 300 teachers. With fewer groups (e.g., two treatments and one control), the total sample size would be 225.

Note that for those wishing to randomly assign schools or grade-level teams, the sample size requirements would be steeper. Although there may ultimately be some benefit to delivering professional development to intact groups of teachers (e.g., school or grade-level teams) because it can help facilitate learning and cooperation, we argue it is not necessary in this early design phase, where the goal is to simplify to compare a variety of program design features to one another.

Although the proposed sample size may seem large in comparison to the standard practice of working with one moderate-sized group of teachers over a very extended period of time, there are several design features that help either mitigate cost or argue for the increased burden. First, as noted above, the professional development groups need not be run concurrently by the same facilitator; in fact, a strength of this design is that multiple facilitators can be hired, trained, and tracked—with the benefit that results do not hinge on the capacity of any single facilitator (see Heller et al., 2012, for an example). To facilitate logistics, group participation in the professional development could be staggered, with half the groups conducted during one academic year and half during the next academic year. Second, we recommend limiting professional development to one school year or less; working with teachers for shorter amounts of time is more reflective of both modal professional development delivery, and of the current situation in schools, where high attrition rates (up to 10% per year from schools according to Schools and Staffing Survey SASS [Kaiser, 2011], and 20%–35% per year in two recent studies we have performed), mean that longitudinal studies have become difficult to sustain.

Furthermore, we believe Stage 2 studies could easily be incorporated into existing professional development experiences or ongoing studies. For example, many professional development experiences take place within a single school district and are designed to provide professional development to hundreds of teachers. In such instances, teachers within the district could be randomly assigned to various permutations of the professional development at very little cost. Similarly, many professional development program providers disseminate their programs to multiple districts before they have been formally evaluated through an efficacy trial. Teachers within these districts could be randomly assigned to different versions of the professional development. Finally, in some instances, such "mini" randomized studies could be embedded within large-scale efficacy trials, which often randomize schools rather than teachers. As such, they generally involve hundreds of teachers who could be randomly assigned to various professional development permutations. By thinking creatively about ways to embed this type of exploratory study into professional development structures that already exist, the field could learn a greater deal about the most effective ways to deliver professional development to teachers.

Examples of Stage 2 and Stage 5

Several recent studies have taken a Stage 2 approach, although often with a much-simplified design. For instance, Russell, Carey, Kleiman, and Venable (2009) randomly assigned middle school algebra teachers to online professional development with varied amounts of support from instructors. Results from this study found that there was no significant difference in changes to pedagogical beliefs among the teachers assigned to the different support conditions, nor was a significant difference found in teachers' instructional practices among the groups. However, here the comparisons between outcomes were made based on a single group of teachers in each condition, rather than multiple groups, making it difficult to disentangle group or facilitator effects from the variations in the treatment itself.

Another example of a Stage 2–type study by Heller et al. (2012) permuted not format, but content—meaning the actual activities teachers engage. Nevertheless, the study design is similar to the one proposed here and instructive for our proposed line of research. The authors kept constant the topic of professional development (electric circuits) but used random assignment to examine the effect of three delivery methods on teaching and learning outcomes. In *Teaching Cases*, teachers read and discussed cases of science teaching written by practicing teachers; in *Looking at Student Work*, teachers analyzed their own students' science productions; in *Metacognitive Analyses*, teachers reflected on their own work. The professional development was not extensive—a total of eight 3-hour sessions—but each condition was delivered multiple times each at eight geographic sites by 12 rotating facilitator pairs, improving the generalizability of findings. Although all three conditions improved students' standardized assessment performance beyond that of a control group, only the first two improved students' written justifications for answers.

In addition, a review of research suggested that in at least two areas, there may be sufficient studies to provide an example of Stage 5 of the proposed approach—that of collecting evidence across multiple studies to understand the impact of different professional development design features on program outcomes. Broad-scale examinations have already been conducted by Kennedy (1999), Scher and O'Reilly (2009), and Yoon et al. (2007), who found that professional development focused on how students learn, professional development focused on both content pedagogy, and more extended programs had substantial effects on student outcomes. However, we also suggest that meta-analyses reviewing Stage 2 work could focus on practical design considerations facing contemporary professional developers as well. To illustrate how this might occur, we conducted searches on terms such as *professional development, control, treatment*, and *random*, in engines such as Education Abstracts through EBSCO and Education Resources Information Center. Articles that contained some or all of these search terms were examined for inclusion. Furthermore, references in reviewed articles that met the criteria were also reviewed, even if those references did not randomize teachers to condition. We found two areas with sufficient articles to conduct an impressionistic review: online versus in-person professional development and analysis of subjects' own versus other subjects' or experts' videotaped practice. In both cases, we found mostly simple-design studies, similar to the Russell study described above; nevertheless, the process of reviewing and synthesizing is instructive, for it would be identical in a real Stage 5 situation.

Online Versus In-Person Professional Development

One issue arising from current professional development practice is whether programs can be delivered equally effectively in-person and online. Although the body of research that examines this contrast is limited (Dede at al., 2009), and most studies solely compare online professional development to a control group that receives no professional development (e.g., Masters, De Kramer, O'Dwyer, Dash, & Russell, 2010; O'Dwyer et al., 2010), the studies that do compare delivery mode holding the program constant can provide a sense for whether there is a mode effect on program outcomes.

Fisher, Schumaker, Culbertson, and Deshler (2010) explored the question of whether virtual and in-person workshops would vary in their effect on teacher learning and teachers' use of their learning from a professional development intervention. In this study, researchers randomly assigned teachers enrolled in a special education course to either the online or in-person setting. In each setting, teachers were provided with the same content materials focused on student learning. Although teachers assigned to the in-person setting reported higher satisfaction levels, there were no significant differences in teacher learning results between the two groups.

Powell et al. (2010) conducted a randomized controlled trial of online versus in-person professional development based on expert literacy coaching. The study focused on 88 classrooms in 24 Head Start centers with a goal of improving evidence-based literacy instruction. Content coverage, facilitators, and time spent on content areas were the same across treatment conditions, where the treatment conditions were in-person versus online expert coaching. Results from this study were mixed. In particular, based on observations of teachers' classrooms, teachers in the in-person coaching condition had larger gains in some areas of instruction compared to teachers in the remote coaching condition. With regard to student outcomes, those with teachers in the online condition showed larger gains in language skills assessed through the Peabody Picture Vocabulary Test. Powell and Diamond (2011) conclude in a later paper that there are not "consistent effects" of delivery mode on student learning outcomes.

Finally, Fishman et al. (2013) examined whether in-person (n = 24) and online (n = 25) professional development around a new science curriculum resulted in different teaching and learning outcomes. Although the in-person professional development was a week-long workshop and the online professional development was asynchronous and could be completed at any time, the topics were the same in both conditions. The researchers concluded that in both conditions, teachers increased their confidence to use the new curriculum materials and used the materials in ways intended by the designers; there were no appreciable differences in student learning between the two treatments.

Although the number of rigorous studies in this domain is small, these early results are suggestive. Aside from more satisfaction with in-person professional development, there were no effects of in-person or online delivery mode on outcomes. Additional evidence of the equivalence between online and in-person professional development would have important implications for school districts as they search for budget efficiencies as well as ways to serve teachers isolated by geography or by the lack of peers in their schools.

Analyzing Teachers' Own Versus Other Teachers' Lessons

Another issue arising from emergent professional development practice is whether teachers should watch and analyze video of their own teaching or stock footage from a library. Although the body of research featuring such comparisons is limited (Seidel, Stürmer, Blomberg, Kobarg, & Schwindt, 2011), Sherin and Han (2004) point to the need for this type of exploration.

Zhang, Lundeberg, Koehler, and Eberhardt (2011) explored the differential effects of video in the context of professional development. Using the *Problem-Based Learning* approach for guiding analysis, science teachers were asked to view and analyze three types of videos of instruction—published, their own, and their peers'. The teachers rated all types of videos as useful for reflecting on their own practices but rated their own as the most useful and the published videos as the least useful.

Seidel et al. (2011) conducted a randomized trial, assigning science teachers to two different conditions—one in which they viewed others' lessons, the other in which they watched their own lessons. At the beginning of the study, all teachers attended a 1-day workshop to learn about using video to reflect on teaching, with the workshops having the same structure and content for both treatment groups. Teachers assigned to the own-video treatment reported feeling more immersed while watching their lessons. Moreover, the researchers reported that analyzing their own lessons provided teachers with a more stimulating experience. However, there was no discernable difference in what teachers noticed across the treatment groups, and the researchers observed that teachers in the own-video condition were less critical and identified fewer consequences of their teaching on student learning.

In a different context, that of individuals learning therapeutic techniques, Baum and Gray (1992) randomly assigned students to four video-watching conditions, one of which consisted of watching an expert therapist in consultation with a client and another of which consisted of watching videotape of the students' own consultation with a client. Students' therapeutic skills were measured before and after this training, with mixed results. Students who observed experts' tapes had the greatest improvements in skill acquisition, whereas the students who watched their own interviews showed the least improvement on objective measures. However, the self-observation group reported the greatest level of satisfaction with their training and gave the most positive feedback.

Although we cannot reach firm conclusions based on this limited set of studies, they suggest that although individuals may prefer to watch videotapes of their own practice, watching videotapes of expert teaching may provide greater benefits to knowledge and skills. Whatever future studies unearth, this example demonstrates that multiple studies of different professional development content can nevertheless be analyzed to discern larger lessons for the professional development community.

Conclusion

In this article, we have proposed a new approach to professional development research and provided examples demonstrating how that approach might work. In this approach, we recommend instituting a series of small-scale but rigorous trials to determine the effects of various program delivery features on mostly proximal outcomes. These trials, which would be conducted at multiple sites with small cadres of developers/facilitators, would both inform future larger scale trials of the specific intervention and also provide advice to local developers regarding the designs most likely to affect teachers. We think this approach would be well-suited to I3 (Investment in Innovation) development-level and IES Goal 2 studies, both of which explicitly ask developers to determine promising practices, although in the latter case, program requirements regarding the use of student outcomes as indicators of success would need to be relaxed.

Clearly, we could have crafted this proposal for moving professional development research forward quite differently. We could have advocated for a more lengthy exploration and development stage (Stage 1) on the view that robust professional development scarcely exists and will take time to produce; we could have argued for nonexperimental rather than experimental methods, or for permuting the content of the program rather than delivery method. Instead, however, we crafted this proposal to lead as far away as possible from current practices—small-scale research, studies with a lengthy development stage, cluster-randomized trials of single programs—in hopes that we would evoke debate and discussion. We are not convinced we are correct and hope that others are not either.

In addition, the field should consider the concrete drawbacks of this proposal. For example, more teachers and sites would need to be enrolled at an earlier stage of the programs' progression, which means greater coordination and, potentially, expense. Studies that rigorously examine program impact on student outcomes would only occur in Stage 3, several years into program development. And research that allows the comparison of program content—for instance, whether coaching or lesson study is more effective and efficient for improving teaching and learning—would still be necessary, requiring other lines of research.

However, we believe that some of the more troublesome issues may be mitigated. The conduct of early-stage professional development across multiple sites, for instance, can be eased by the hiring and training of several part-time facilitators (e.g., retired teachers or coaches). Experimenting with program delivery method on a large number of groups of teachers is also warranted given that this, we suspect, is how most professional development is implemented—at large scale, even in its initial stages.

We also believe that the benefits may outweigh the costs of this approach. In a world in which the success of programs may be largely driven by implementation considerations that are in turn influenced by variability in local contexts (Penuel, Fishman, Cheng, & Sabelli, 2011), it is important to understand whether programs are equally effective across sites. It is also important to test programs rigorously prior to the final design, such that the most effective approach can be identified. Finally, the production of usable knowledge about program design for local practitioners would be a strong contribution in the next era of professional development research.

NOTES

The authors would like to thank Hilda Borko, Barry Fishman, Mike Garet, Kirk Walters, and three anonymous reviewers for helpful comments on an earlier draft. Work for this article has been funded by the NSF DRK-12 program (1221693) as well as the National Center for Teacher Effectiveness (R305C090023).

[1]For example, random assignment could take place within schools with a quarter of the teachers in each school randomly assigned to each of the four conditions.

[2]At this stage of program development, we would argue for not administering supplemental assessments to students; doing so raises the expense of the study considerably.

[3]We relax the required significance level as we are somewhat less concerned about a false conclusion that a particular permutation was more effective than another than we would be if we were assessing the program's overall effectiveness.

REFERENCES

Angrist, J. D., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics*, *19*, 343–369.

Arens, S. A., Stoker, G., Barker, J., Shebby, S., Wang, X., Cicchinelli, L. F., & Willams, M. J. (2012). *Effects of curriculum and teacher professional development on the language proficiency of elementary English language learner students in the central region* (NCEE 2012-4013). Denver, CO: Mid-Continent Research for Education and Learning.

Baum, B. E., & Gray, J. J. (1992). Expert modeling, self-observation using videotape, and acquisition of basic therapy skills. *Professional Psychology: Research and Practice*, *23*, 220–225.

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

Borko, H., Stecher, B. M., Alonzo, A. C., Moncure, S., & McClam, S. (2005). Artifact packages for characterizing classroom practice: A pilot study. *Educational Assessment*, *10*, 73–104.

Bos, J. M., Sanchez, R. C., Tseng, F., Rayyes, N., Ortiz, L., & Sinicrope, C. (2012). *Evaluation of Quality Teaching for English Learners (QTEL) professional development* (NCEE 2012-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bush, W. S., Ronau, R., Brown, T. E., & Myers, M. H. (2006, April). *Reliability and validity of diagnostic mathematics assessments for middle school teachers*. Paper presented at the American Educational Association Annual Meeting, San Francisco, CA.

Buysse, V., Castro, D. C., & Peisner-Feinberg, E. (2010). Effects of a professional development program on classroom practices and outcomes for Latino dual language learners. *Early Childhood Research Quarterly*, *25*, 194–206.

Cabalo, J. V., Ma, B., & Jaciw, A. (2007). *Comparative effectiveness of professional development and support tools for world language instruction: A report on a randomized experiment in Delaware*. Palo Alto, CA: Empirical Education Inc.

Carlisle, J. F., Kelcey, B., Rowan, B., & Phelps, G. (2011). Teachers' knowledge about early reading: Effects on students' gains in reading achievement. *Journal of Research on Educational Effectiveness*, *4*, 289–321.

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

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

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science Magazine*, *315*(5811), 464–465.

Daehler, K. R., & Shinohara, M. (2001). A complete circuit is a complete circle: Exploring the potential of case materials and methods to develop teachers' content knowledge and pedagogical content knowledge of science. *Research in Science Education*, *31*, 267–288.

Dede, C., Ketelhut, D. J., Whitehouse, P., Breit, L., & McCloskey, E. M. (2009). A research agenda for online teacher professional development. *Journal of Teacher Education*, *60*, 8–19.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, 181–199.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, *66*, 1–27.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher*, *31*(8), 4–14.

Finklestein, N., Hansen, T., Huang, C.- W., Hirschman, B., & Huang, M. (2011). *The effect of problem based economics on high school economics instruction* (NCEE 2010-4002). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Fisher, J. B., Schumaker, J. B., Culbertson, J., & Deshler, D. D. (2010). Effects of a computerized professional development program on teacher and student outcomes. *Journal of Teacher Education*, *61*, 301–312.

Fishman, B., Konstantopoulos, S., Kubitskey, B. W., Vath, R., Park, G., Johnson, H., & Edelson, D. C. (2013). Comparing the impact of online and face-to-face professional development in the context of curriculum implementation. *Journal of Teacher Education*, *64*(5), 426–438. doi:10.1177/0022487113494413

Frechtling, J. A., Sharp, L., Carey, N., & Vaden-Kiernan, N. (1995). *Teacher enhancement programs: A perspective on the last four decades*. Retrieved January 11, 2013, from http://www.physics.ohio-state.edu/~jossem/REF/151.pdf

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., . . . & Sztejnberg, L. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, 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*(4), 915–945.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., . . . & Doolittle, F. (2011). *Middle School Mathematics Professional Development Impact Study: Findings after the second year of implementation* (NCEE 2011-4025). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal*, *47*, 694–739.

Good, T., Grouws, D., & Ebmeier, H. (1983). *Active mathematics teaching*. New York: Longman.

Grossman, P. L., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J. H., Boyd, D. J., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (Working Paper 16015). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*, 798–812.

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.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333–362.

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

Institute for Educational Sciences. (2012). *Request for applications, education research grants* (CFDA Number: 84.305A). Washington, DC: Author.

Jacob, R., Goddard, R., & Kim, E.S. (n.d.). *Assessing the use of aggregate data in the evaluation of school-based interventions: Implications for evaluation research and state policy regarding public use data*. Manuscript submitted for publication.

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, *39*, 50–79.

Kaiser, A. (2011). *Beginning teacher attrition and mobility: Results from the first through third waves of the 2007-08 Beginning Teacher Longitudinal Study* (NCES 2011-318). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Kennedy, M. M. (1999). Form and substance in mathematics and science professional development. *NISE Brief*, *3*(2), 1–7.

Landry, S. H., Anthony, J. L., Swank, P. R., & Monseque-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology*, *101*, 448–465.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*, 25–47.

Masters, J., De Kramer, R. M., O'Dwyer, L. M., Dash, S., & Russell, M. (2010). The effects of online professional development on fourth grade English Language Arts teachers' knowledge and instructional practices. *Journal of Educational Computing Research*, *43*, 355–375.

McCrory, R., Floden, R., Ferrini-Mundy, J., Reckase, M. D., & Senk, S. L. (2012). Knowledge of algebra for teaching: A framework of knowledge and practices. *Journal for Research in Mathematics Education*, *43*, 584–615.

National Research Council. (2002). *Scientific research in education*, R. Shavelson & L. Towne Eds. Washington, DC: Committee on Scientific Principles for Educational Research, National Academy Press.

Newman, D., Finney, P. B., Bell, S., Turner, H., Jaciw, A. P., Zacamy, J. L., & Gould, L. F. (2012). *Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative* (*AMSTI*) (NCEE 2012-4008). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

O'Dwyer, L. M., Master, J., Dash, S., De Kramer, R. M., Humez, A., & Russell, M. (2010). *E-learning for educators: Effects of on-line professional development on teachers and their students: Findings from four randomized trials*. Retrieved July 9, 2010, from http://www.bc.edu/research/intasc/PDF/EFE_Findings2010_Report.pdf

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher*, *40*, 331–337.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, *44*, 921–958.

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: A comparison of three professional development programs. *American Educational Research Journal*, *48*(4), 996–1025.

Powell, D. R., & Diamond, K. E. (2011). Improving the outcomes of coaching-based professional development interventions. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (Vol. 3, pp. 295–307). New York: Guilford.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational Psychology*, *102*, 299.

Ramey, S. L., Ramey, C. T., Crowell, N. A., Grace, C., & Timraz, N. (in press). The dosage of professional development for early childhood professionals: How the amount, density, and duration of professional development may influence its effectiveness. In J. A. Sutterby (Ed.), *Early childhood professional development: Research and practice through the early childhood educator professional development grant*. Boston: The Emerald Group.

Russell, M., Carey, R., Kleiman, G., & Venable, J. D. (2009). Face-to-face and online professional development for mathematics teachers: A comparative study. *Journal of Asynchronous Learning Networks*, *13*, 71–87.

Russell, M., Kleiman, G., Carey, R., & Douglas, J. (2009). Comparing self-paced and cohort-based online courses for teachers. *Journal of Research on Technology in Education*, *41*, 443–466.

Santagata, R., & Angelici, G. (2010). Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. *Journal of Teacher Education*, *61*, 339–349.

Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2011). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. *Journal of Research on Educational Effectiveness*, *4*, 1–24.

Scher, L., & O'Reilly, F. (2009). Professional development for K–12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness*, *2*(3), 209–249.

Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching & Teacher Education*, *27*, 259–267.

Sherin, M. G., & Han, S. Y. (2004). Teacher learning in the context of a video club. *Teaching & Teacher Education*, *20*, 163–183.

Sherin, M. G., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, *60*, 20–37.

Shinohara, M., Daehler, K. R., & Heller, J. L. (2004, April). *Using a pedagogical content framework to determine the content of case-based teacher professional development*. Paper presented at the annual meeting of the National Association of Research in Science Teaching, Vancouver, Canada.

Smith, S. P., & Banilower, E. R. (2006, April). *Measuring teachers' knowledge for teaching force and motion concepts.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Sowder, J. T., Philipp, R. A., Armstrong, B. E., & Schappelle, B. P. (1998). *Middle-grade teachers' mathematical knowledge and its relationship to instruction: A research monograph*. Albany: State University of New York Press.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805.

U.S. Department of Education. (2002). *U.S. Department of Education strategic plan, 2002-2007*. Retrieved June 15, 2012, from http://www.ed.gov/pubs/stratplan2002-07/index.html

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, *37*, 469–479.

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*. Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Zhang, M., Lundeberg, M., Koehler, M. J., & Eberhardt, J. (2011). Understanding affordances and challenges of three types of video for teacher professional development. *Teaching and Teacher Education*, *27*, 454–462.

AUTHORS

HEATHER C. HILL is a professor at the Harvard Graduate School of Education, 6 Appian Way #445; *heather_hill@harvard.edu*. Her research focuses on teacher and teaching quality.

MARY BEISIEGEL, PhD, is an assistant professor at Oregon State University, 368 Kidder Hall, Corvallis, OR 97331; *mary.beisiegel@oregonstate.edu*. Her research focuses on qualities of effective mathematics instruction in K–12 and postsecondary settings.

ROBIN JACOB, PhD, is a research assistant professor at the University of Michigan's Institute for Social Research and the School of Education, 426 Thompson Street, Perry Room 2338, Ann Arbor, MI 48104; *rjacob@umich.edu*. Her research focuses on how policies and programs can affect instructional quality and outcomes in elementary schools.

Manuscript received January 18, 2013
Revisions received May 22, 2013, and September 25, 2013
Accepted September 27, 2013