

# Experimenting With Teacher Professional Development: Motives and Methods

Andrew J. Wayne, Kwang Suk Yoon, Pei Zhu, Stephanie Cronen, and Michael S. Garet

A strong base of research is needed to guide investments in teacher professional development (PD). This article considers the status of research on PD and articulates a particular direction for future work. Little is known about whether PD can have a positive impact on achievement when a program is delivered across a range of typical settings and when its delivery depends on multiple trainers. Despite a consensus in the literature on the features of effective PD, there is limited evidence on the specific features that make a difference for achievement. This article explains the benefits offered by experiments in addressing current research needs and—for those conducting and interpreting such studies—discusses the unique methodological issues encountered when experimental methods are applied to the study of PD.

**Keywords:** experimental design; professional development; research methodology; school/teacher effectiveness; staff development; teacher education/development

Two decades ago, a groundbreaking study was published demonstrating that teacher professional development (PD) could improve student achievement. Carpenter, Fennema, Peterson, Chiang, and Loef (1989) randomly assigned 40 first-grade teachers to two groups. One group received a brief, 4-hour PD program. The other group received an extensive 80-hour program known as cognitively guided instruction (CGI). The students of the teachers who received CGI outperformed the students of the other teachers on three of the six student achievement measures that were examined.

Bolstered by findings like this, policy makers have sought to make the PD that teachers participate in more effective at raising student achievement. For instance, the PD supported under the No Child Left Behind Act (NCLB) is expected to improve the quality of teaching and boost student achievement, and NCLB encourages school districts to adopt programs and practices that are supported by scientifically based research (Birman et al., 2007). Other sponsors of PD initiatives, such as the National Science Foundation, have also sought to improve the quality of PD made available to teachers and have been eager to conduct research on its effects (see, e.g., Blank, de las Alas, & Smith, 2008; Supovitz, Mayer, & Kahle, 2000).

But studies have not yet provided the kind of clear guidance needed to steer investments in PD. In 2004–2005, federal funds spent on PD amounted to approximately \$1.5 billion (Birman et al., 2007). This level of investment—plus the investments made by states and school districts—necessitates a strong base of research to guide policy and practice.

In this article, we consider the current status of research on PD<sup>1</sup> and articulate a particular direction for the future. In the first half of the article, we discuss the research on PD. This research shows that PD, when delivered in conducive settings by those who designed the PD, can have a positive impact on student achievement. But we know much less about whether PD can have a positive impact on achievement when a program is delivered in a range of typical settings and when it is delivered by multiple trainers. In addition, although a consensus has emerged in the literature about the features of effective PD, the evidence on the specific features that make a difference for achievement is weak, and the consensus falls short of addressing several practical questions faced by those who design and fund PD.

During the course of the article, we argue that randomized experiments have a major role to play in future research on PD. In the second half, we discuss experiments and other methodological options, noting that experiments present some distinct advantages when applied to the study of PD. We then discuss some of the unique design challenges encountered in the use of experiments to study the effects of PD. What seem like simple decisions—what PD program to study, where to conduct studies, how to structure the sample, and what measures to include—present several significant challenges in studies of PD.

## Studies of the Influence of Teacher Professional Development

We focus first on what has been learned in studies of the influence of teacher PD on student achievement. Kennedy's (1998) literature review focusing on mathematics and science PD programs was perhaps the first widely circulated review to address this topic. Building on the literature reviews by Kennedy and by Clewell, Campbell, and Perlman (2004), Yoon, Duncan, Lee, Scarloss, and Shapley (2007) recently conducted the most systematic and comprehensive review to date.

Yoon et al. (2007) examined studies of impacts in three core academic subjects (reading, mathematics, and science). They focused the review on studies that met the What Works Clearinghouse (WWC) evidence standards. In total, 9 studies

emerged as meeting the WWC evidence standards from 132 identified as relevant. The 9 studies all focused on elementary school teachers and their students. Five studies were experiments that met evidence standards “without reservations”; the remaining four studies met evidence standards “with reservations” (one experiment with a group equivalence problem and three quasi-experiments).

On one hand, the results of the studies were promising. Pooling across the studies in which effect size was reported in terms of student-level standard deviations, the average overall effect size was .55.<sup>2</sup> This average effect size looks remarkably high when compared with what is found in other studies of the influence of teacher variables on student achievement. For example, in their evaluation of Teach for America (TFA), Decker, Mayer, and Glazerman (2004) randomly assigned students to TFA teachers and to other newly assigned novice teachers. The effect size on students’ mathematics scores was .26 student standard deviations.<sup>3,4</sup>

On the other hand, these studies did not involve PD programs delivered in a variety of settings and led by multiple trainers. Instead, the studies involved a small number of teachers, ranging from 5 to 44, often clustered in a few schools. In addition, the developers of the PD provided it directly to teachers. Studies of this type are sometimes termed *efficacy trials*, in contrast to *effectiveness trials*. Efficacy trials take place under conditions that are conducive to obtaining an effect. In an effectiveness trial, an intervention is tested in the full range of settings in which it is designed to work (see Kellam & Langevin, 2003; Shadish, Cook, & Campbell, 2002; Society for Prevention Research, 2004). Results from an effectiveness trial are more likely to be relevant to those considering the adoption of specific PD programs in a particular school or district.

In sum, one of the major challenges in research on the influence of PD on student achievement is to determine whether PD programs can be effective when delivered in typical settings by those not involved in the development of the PD programs. This is a logical step in the progression of research; it is what Borko (2004) called Phase 3 studies of PD in her presidential address to the American Educational Research Association in 2004. She recommended that researchers continue studying teacher PD programs and commended their efforts, she also articulated a three-phase pipeline of research. The pipeline culminates in studies showing that particular PD programs could be adopted in a range of settings, with consistent effects on teaching and learning.<sup>5</sup>

### **The Features That Make Professional Development Effective**

In addition to showing that PD could be effective, Kennedy’s (1998) review sought to identify the features of effective PD programs. To do so, Kennedy categorized studies according to the PD being studied. She found that the relevance of the content of the PD was particularly important. She classified in-service programs into four groups according to the level of prescriptiveness and the specificity of the content they provide to teachers.<sup>6</sup> On the basis of her analysis of effect sizes, Kennedy concluded, “Programs whose content focused mainly on teachers’ behaviors demonstrated smaller influences on student learning than did

programs whose content focused on teachers’ knowledge of the subject, on the curriculum, or on how students learn the subject” (p. 18). Kennedy’s literature review suggested an important role for content emphasis in high-quality and effective PD. Her seminal work prompted others to test the same research hypothesis in their subsequent studies (cf. Desimone, Porter, Garet, Yoon, & Birman, 2002; Garet, Porter, Desimone, Birman, & Yoon, 2001; Yoon, Garet, Birman, & Jacobson, 2006).

In the recent Yoon et al. (2007) review, there was relatively little variation in the features of the PD in the nine studies that met the evidence standards for inclusion in the review, and thus the authors were unable to draw strong conclusions about the features of PD programs that make them effective.

Despite the lack of solid evidence, drawing on various bodies of theory and correlational and case study evidence, a consensus has been built on promising “best practices” (Garet et al., 2001; Guskey, 2003; Hawley & Valli, 1998; Kennedy, 1998; Little, 1993; Loucks-Horsley, Hewson, Love, & Stiles, 1998; National Commission on Teaching and America’s Future, 1996; Showers, Joyce, & Bennett, 1987; Wilson & Berne, 1999). For example, it is generally accepted that intensive, sustained, job-embedded PD focused on the content of the subject that teachers teach is more likely to improve teacher knowledge, classroom instruction, and student achievement. Furthermore, active learning, coherence, and collective participation have also been suggested to be promising best practices in PD (Garet et al., 2001).

It is important to recognize that this consensus—although it has endured for more than a decade—lacks sufficient specificity to guide practice. For example, nearly everyone decries the “one shot” workshop and affirms that PD should be “sustained” and “intensive.” And among the studies identified by Yoon et al. (2007), there is at least suggestive evidence that PD is more likely to be effective when given in larger “doses.”<sup>7</sup> But the cost of developing and delivering PD grows proportionally with the number of days involved, and requiring teachers to be out of the classroom on regular school days is disruptive to student learning. More rigorous research designs are needed to resolve these dilemmas—by determining the relative effectiveness of PD programs with different durations or different allocations of PD events across time.

Another example of the need for greater specificity to guide practice is the consensus that PD should be “school based” or “integrated into the daily work of teachers” (see Hawley & Valli, 1998; Joyce & Showers, 2002). Such PD typically requires that a coach or mentor work with teachers at one or more schools, which is among the most expensive approaches to PD available. With what frequency, duration, and quality would coaching or mentoring need to occur to make a difference? And suppose the budget is fixed. Should the amount of off-site PD be reduced in order to increase the amount of school-based PD? These are simple, practical questions faced by those who design and fund PD initiatives.

### **Designs to Address Questions About Professional Development**

Thus far, we have argued that future research on PD should examine two main questions: (a) whether PD programs implemented by study authors or their colleagues, in a conducive context, remain

effective when delivered by others, in more typical contexts, and (2) what specific features of PD matter—for example, how much PD is enough to ensure effectiveness and whether adding school-based components to PD programs is worthwhile.

Addressing questions like those posed above requires study designs that allow causal inferences. Researchers studying the effects of PD have made claims of causality using several designs—the most persuasive of which have been experimental and quasi-experimental designs. (The nine studies that became the focus of the Yoon et al., 2007, review were experiments and quasi-experiments.) Under some conditions, observational studies can also be reasonably persuasive. In this section, we discuss the strengths and weaknesses of these three approaches to answering questions about the impact of PD.

Observational studies acquire their name from the fact that they involve observation only. Such studies observe and analyze natural variation, which for our purposes includes variation in at least PD and student achievement. They usually use administrative data or data from government-funded survey programs. For relatively little cost, they can seek relationships in large samples, encompassing a variety of contexts. Harris and Sass (2007), for example, used administrative data from Florida on teachers' participation in PD; the measures of PD were the number of hours of content-oriented PD and the number of hours of other PD. They tested whether certain measures of PD were related to teacher effectiveness, as measured by teachers' ability to raise student achievement. Owing to the size of the Florida database, the researchers were able to examine the relationship between PD and teacher effectiveness in both reading and mathematics, and they further disaggregated their results by elementary, middle, and high school.

Experiments and quasi-experiments go beyond observation to include manipulation of the treatment—that is, the PD (cf. Shadish et al., 2002). The study determines the selection of sites and defines who will be eligible to receive the treatment. Importantly, the study also specifies what the PD will be, so the findings usually pertain to a particular PD program.

Another important aspect of the three designs is the handling of selection bias—the problem of having participants with different characteristics in the treatment and comparison groups. We cannot assume that teachers who typically receive PD are equivalent in every way to teachers who do not. Much PD is voluntary, and teachers who volunteer to participate may differ in motivation or prior knowledge and instructional practice from teachers who do not; other PD is mandated for schools identified for improvement under accountability systems (Birman et al., 2007), and teachers teaching in such schools may differ from teachers in other schools not targeted for PD.

Experiments deal with this problem by formatting a treatment group and a control group are formed through random assignment. Random assignment, as such, does not create any bias. Importantly, random assignment ignores both the measured characteristics of those being assigned and their unmeasured characteristics.

Observational studies and quasi-experiments cannot eliminate the possibility of selection bias, but they can reduce it, especially insofar as pertinent data are measured. In observational studies,

it is common to minimize selection bias by estimating the treatment effect in a model that includes other teacher-linked factors that may correlate with both participation in PD and the student outcome measure. For example, it is generally accepted that students learn more from teachers who have a degree in their subject. Having a degree in one's subject may also be correlated with one's participation in PD.

Another approach in observational studies to minimize selection bias is to use information on the process by which teachers are assigned either to receive the treatment or to not receive it. For example, using a technique known as selection correction (Heckman, 1979), one could model the likelihood that a teacher with particular measured characteristics attends the PD. That information can be incorporated into an overall model to develop a less-biased estimate of the effect of the PD.

For quasi-experiments, the same techniques are available for minimizing selection bias during analysis.<sup>8</sup> However, quasi-experiments generally allow one to define the comparison group; therefore, quasi-experiments often involve matching. Through matching, one tries to ensure that the comparison group is equivalent to the treatment group by selecting comparison schools with the same measured characteristics as the treatment schools. For example, McCutchen et al. (2002) recruited teachers from a large metropolitan area. Their study identified pairs of schools with similar poverty levels. One school from each pair was designated to receive a 100-hour program of PD, and the other was designated to serve as a control. The authors chose not to make these designations on the basis of random assignment. Instead, to capitalize on the potential benefits of having multiple teachers from one school learning together as a group, the authors favored designating as treatment schools those schools where multiple teachers had volunteered. The study is therefore considered a quasi-experiment with matching. The authors' data showed that the teachers and students in the treatment and comparison groups started with similar characteristics at baseline.

Note that matching does not rule out the possibility of some residual selection bias; the matching could have left some differences in variables that were not measured. In general, the techniques described above for minimizing selection bias work only insofar as the necessary data are measured (Shadish et al., 2002).<sup>9</sup>

In sum, experiments, quasi-experiments, and observational studies each have different advantages. Observational studies are likely to continue to be an important source of suggestive findings, which guide investments in experiments and quasi-experiments. Experiments and quasi-experiments provide a means to build the research base by testing specific interventions. Experiments have the advantage of eliminating selection bias through random assignment; selection bias is a significant concern in studies of PD. However, experiments are not feasible in all contexts; various practical concerns may prevent random assignment. In such situations, well-designed quasi-experiments are a useful alternative.

### Issues to Consider When Designing Experiments and Quasi-Experiments

We now turn to a discussion of the methodological issues faced by those who design experiments—and to some extent by those who design quasi-experiments. Researchers designing randomized

controlled trials focused on teacher PD interventions face a common set of methodological issues. In this section, we discuss the trade-offs inherent in each issue. The resolution of these issues depends in part on the particular intervention selected and the resources available. We organize the discussion under five broad subheadings.

### *Design Issue 1: What Treatment(s) Will Be Studied?*

The motive for selecting a particular treatment for study is usually the expansion of the knowledge base. Given evidence that certain PD interventions affect student achievement, one may want to test a competing PD intervention or a promising variant of an intervention that has already been tested.

One challenge in attempting to expand the knowledge base this way is that a PD intervention rests on at least two theories, which we will call the *theory of instruction* and the *theory of teacher change* (for a related argument, see Supovitz, 2001). The theory of instruction is the intervention's theory about the links between the specific kinds of teacher knowledge and instruction emphasized in the PD and the expected changes in student achievement. For example, one PD intervention may be predicated on the theory (of instruction) that students gain more proficiency in reading when taught using a phonics-based approach. The theory of teacher change is the intervention's theory about the features of PD that will promote change in teacher knowledge and/or teacher practice, including its theory about the assumed mechanisms through which features of the PD are expected to support teacher learning. The theory of teacher change is not limited to the structural features of the PD, such as its duration and span, but also includes elements of the activities in which teachers are expected to engage during the PD (e.g., opportunities for teachers to review student work) and the intermediate teacher outcomes these activities are expected to foster.

In contrast to the near consensus on the features of PD worth testing, there is much less consensus on the theories of instruction. The PD interventions tested in recent studies differ widely in the theory of instruction being tested. Some interventions focus on a specific set of instructional practices. For example, Sloan (1993) randomly assigned a treatment that sought to elicit teaching behaviors associated with the direct instruction model—specifically, Madeline Hunter's "seven steps of the teaching act." Participating elementary teachers were expected to use these practices in teaching all subjects.

Other PD interventions use a theory of instruction that places greater focus on building a teacher's knowledge of a content area or of student thinking (see, e.g., Carpenter et al., 1989), with the expectation that this increase in knowledge will lead to improvements in the quality of teaching more generally. These PD interventions are less prescriptive with respect to instruction but still embed a theory of instruction—one that includes decision making on the part of the teacher.

Because any given PD intervention requires these two theories, studies of PD interventions are tests of a package—a package that inevitably draws on both a theory of teacher change and a theory of instruction. This reality makes the task of building the knowledge base more difficult. For example, if Sloan (1993) had found no effect, we would not know whether the flaw was the direct instruction model or the teacher change model. Structurally, the PD was a

30-hour treatment that included summer sessions and seven follow-up meetings, spread out over a period of 6 months.<sup>10</sup>

Researchers interested in "unpacking" the packages of PD—that is, unconfounding the theory of instruction and the theory of teacher change—may want to specify two versions of their PD. For example, one study currently under way—the Institute of Education Sciences' Study of the Impact of Two Professional Development Interventions in Early Reading—randomly assigns schools to a control group or one of two treatment groups.<sup>11</sup> Both treatments include the same core PD program of summer training and school-year seminars, but the second treatment adds a coaching component provided by part-time coaches based at each school. The study will show whether or not the coaching component makes a difference.

Such "three-arm" studies are a promising way to add to the knowledge base because one can test the importance of specific PD components while holding constant the theory of instruction.<sup>12</sup> The most serious constraint in specifying the treatment conditions for such three-arm studies is that the PD received by the two treatment groups needs to differ enough to result in measurable differences in student achievement (see Issue 4, below).

### *Design Issue 2: In What Contexts Will the Professional Development Be Studied?*

In describing the existing literature, we have noted the need to demonstrate the effectiveness of PD interventions when delivered in typical contexts by different persons. That is most appropriate when some prior evidence of efficacy is available for a particular PD intervention (Carpenter et al., 1989; Good, Grouws, & Ebmeier, 1983; Saxe, Gearhart, & Nasir, 2001). Given evidence that an intervention is effective in the context of a high-functioning district or set of schools, it is important to determine whether the intervention can remain effective when delivered in other settings, such as under-resourced schools, or whether the intervention can be delivered by persons other than its creators without loss of effectiveness.

Regardless of who delivers the PD and the variety of contexts incorporated, it is important to realize that selecting the specific schools and/or districts requires some review of the fit between the PD and the features of those locations. In this section, we consider some additional issues involved in selecting the context in which the PD will be studied, and we discuss the need to measure the PD received by all teachers during the study.

#### *Curricular Context*

First, it is necessary to consider the curricular context and find locations with curricula for which the PD is suited. Curricula, like PD programs, embed specific theories of instruction. Ideally, it is sensible to seek locations that use curricula that align with the underlying theory of instruction embedded in the PD. At the very least, it is necessary to choose locations where the curricula would not discourage the practices promoted by the PD.<sup>13</sup>

#### *Ambient Professional Development*

Second, it is important to examine the PD that already exists in districts being considered. It is inevitable that during the study, the teachers in both the treatment and the control groups will receive other PD. Teachers participate in a variety of PD activities each year

because of mandates, incentives, or personal initiative (see Choy, Chen, & Bugarin, 2006). Teachers may be part of informal groups at their schools that serve PD needs. Teachers will presumably continue to participate in all these PD experiences regardless of their status in the study, except to the extent researchers are able to negotiate special arrangements. We refer to this PD as the *ambient PD*, to indicate that it pervades the context in which the study takes place.

The existence of ambient PD is problematic because an experiment measures impact as the difference in outcomes between the treatment group and the control group. To the extent that the content of the ambient PD overlaps with the content of the study PD, the difference in outcomes between the treatment and the control conditions may be reduced. For instance, suppose researchers wanted to study a 10-day program of content-focused PD for geometry teachers. If the district in which the study took place provided all geometry teachers with 2 days of workshops on geometry, the teachers in the treatment group and the control group would already have experienced some of what the 10-day program addressed. One way to address such a problem is to conduct one's study in districts where the content of the study PD is least likely to overlap with the content of the ambient PD; another is to increase the intensity of the study PD to sharpen the contrast with the ambient PD.

Through selection of the study context, one can also avoid selecting contexts in which the ambient PD would contradict the study PD. For instance, the ambient PD might promote use of district-approved instructional materials, whereas the study PD promotes use of other materials.

Another potential concern with ambient PD is whether teachers can attend both the ambient PD and the study PD. Ideally, the teachers in the treatment group will receive all the study PD and will continue to receive the ambient PD, such that the study will show the value added of the PD treatment that is being randomly assigned. But if the treatment is time intensive or if a district has a lot of ambient PD, scheduling conflicts could occur or teachers could begin to feel overloaded and selectively not attend some PD events. Thus treatment group teachers might not get some PD that they were supposed to get. The treatment-control difference would be attenuated because the treatment group would be receiving less of the ambient PD than the control group receives. Alternatively, treatment group teachers might decide to attend the ambient PD in lieu of attending the study PD.

Thus it is important to select locations where the chance of interference with the PD under study is low; alternatively, it may be wise to construct treatments that can be delivered without interfering significantly with other PD. It is also best to find locations where the ambient PD does not share elements in common with the study PD.

#### *Measuring PD Experiences*

Although careful site selection can minimize problematic forms of ambient PD, it is not realistic to assume that site selection alone will resolve all problems. The remaining problems can be addressed in part by measuring PD experiences.

First, it is always possible that the study PD is not delivered or received as intended. One can measure the extent to which the treatment PD was delivered as intended. Doing so may require

complex measures, which ideally would be completed by trained observers. A more basic measure of the delivery of the treatment PD is teacher participation; it is clearly important to document whether some treatment teachers missed large proportions of the treatment PD.

Because those participating in the treatment PD have less time to participate in other PD, it is important to document the service contrast, that is, the treatment-control difference in PD experiences. This information can be collected through surveys administered to both treatment and control teachers, designed to examine the duration and features of teachers' PD experiences over the period of the study.

#### *Design Issue 3: At What Level Should Random Assignment Take Place?*

In theory, a study could evaluate the impact of a PD intervention using district-level random assignment, meaning that each participating district would be assigned by lottery to either receive the treatment or not receive the treatment. Studies of PD interventions have instead randomized at the school or teacher level because these designs use fewer resources yet allow one to ask the same questions. Under teacher-level assignment, individual teachers in each school are randomly assigned to either the treatment or the control condition; under school-level random assignment, the entire group of teachers at each participating school is randomly assigned to either the treatment or the control condition (cf. Donner & Klar, 2000; Murray, 1998). In this section, we frame the choice of teacher- or school-level random assignment in terms of three issues: teacher mobility, spatial concentration, and relative sample size.

#### *The Treatment of Teacher Mobility*

The choice of unit of assignment has important implications for the set of teachers and their students for whom impact is examined. If the teacher is selected as the unit, then each *teacher* included in the study would be assigned to treatment or control at the start of the study, and the teacher (and the teacher's students) would be followed during each wave of data collection. If, however, the school is selected as the unit, then each *school* included in the study would be assigned to treatment or control at the start of the study, and all relevant teachers teaching in each school (and their students) would be the target of data collection at each wave.

Therefore, in the presence of teacher turnover, the two choices for the unit of assignment lead to different teacher samples over time. The teacher-level design involves following teachers as long as they remain in teaching, and collecting data on them and on their students' achievement at the time points specified in the design. Teachers who exit teaching or who change grade levels or subjects must be dropped from the study because data on classroom instruction and student achievement could not be collected from such teachers, even if the teachers could be located. Thus in the teacher-level design, if the PD treatment affects teacher turnover rates, mobility could bias the results by affecting who remains in the treatment group (i.e., selection bias), and this bias would need to be taken into account in the analysis.

In the school-level design, if a teacher leaves a school in the sample over the course of the study and another teacher enters,

then the new replacement teacher would be included in the study sample from that point forward. Thus in a school-level design, mobility can cause the impact of the PD to be diluted. For example, if a treatment teacher left in the middle of a program year, after the PD treatment was complete, and his or her class was handed over to a new teacher who had not been exposed to the PD intervention, then the amount of exposure to treatment that the students in this class had would be cut in half. The treatment impact estimated on the basis of the achievement data from this class of students would reflect the program impact from half of the intended treatment, and it would likely be different from what the students would have experienced had the treatment teacher stayed for the whole program year. This suggests that in a school-level design, it may be desirable to incorporate components in the intervention to provide support for new teachers who enter treatment schools over the course of the study (i.e., supplemental PD).<sup>14</sup>

When a school-level design is used, it is also possible for a teacher in a treatment school to move to a control school (or vice versa), resulting in crossover cases (cf., Shadish et al., 2002). Detailed monitoring of teacher movements is necessary so that the extent of crossover can be documented. In studies involving strong partnerships, the parties may be willing to minimize transfers between treatment and control schools.

#### *The Role of Spatial Separation*

To choose between the school and the teacher as the unit of random assignment, one must consider the value of separating the treatment and control teachers spatially.

First, one threat to the methodological integrity of a random-assignment research design is the possibility that some control group members will be exposed to the program, thus reducing the service contrast—the difference between the group receiving the intervention (or the treatment group) and the “business as usual” group (or the control group). Such contamination of the control group is possible in the case of PD interventions. Because many of these interventions incorporate collaboration among teachers at a given grade level, if some teachers in a school are randomly assigned to an intervention, they are likely to share some of the information provided through the intervention with peers who have been randomly assigned to the control group. This second-hand exposure will attenuate the service contrast and make it difficult to interpret impact estimates. Such exposure is especially likely in schools where teachers collaborate regularly or team teach. By randomly assigning schools to treatment condition, one separates the treatment group from the control group spatially and hence blocks some potential paths of information flow and reduces control group contamination.

Using the school as the unit of assignment may also help deliver the services more effectively by capitalizing on economies of spatial concentration. Spatial concentration of the target group members may reduce the time and money costs of transportation to and from the program and may enable staff members to operate the program more effectively by exposing them to the contexts in which the teachers operate. For example, when a PD intervention involves intensive coaching, it certainly reduces costs if a coach needs to travel to one school to work with four teachers in that school rather than travel to two different schools to

work with two teachers in each of them. School-level randomization enables the coach to spend more time in the school and get exposure to the common problems in that school and thus adjust the delivery of service to fit the school needs more effectively.

Another very different reason for randomly assigning schools is to facilitate political acceptance of randomization by avoiding situations that might appear unfair to participants. Even though random assignment of teachers treats individual teachers equally in the sense that each one of them has an equal chance of being offered the program, this fact is often overlooked; after randomization, treatment group teachers have access to the intervention, whereas control group teachers do not. This perception is especially acute if within a school some teachers receive the intervention and some do not. Therefore, school-level randomization is generally easier to “sell” than teacher-level randomization.

#### *Relative Sample Size*

However, using teachers as the unit of assignment requires fewer participating schools and teachers for a given level of precision. It is well documented that estimates based on cluster randomization have less statistical precision than those based on individual randomization because possible correlations of impacts across individuals within a cluster have to be accounted for in cluster randomization (Bloom, 2005; Murray et al., 1994). By analogy, for a given set of data, estimates based on school-level (cluster) randomization have less statistical precision than those based on teacher-level (individual) randomization because possible correlations of impacts across teachers in a school have to be accounted for if schools are being randomized.<sup>15</sup> In other words, it requires fewer teachers to detect an impact of a given size with a given level of precision if the randomization is done at the teacher level instead of at the school level. Note, however, that the difference in precision between these two options depends on specific outcomes and analytical methods used in the evaluation and can vary from program to program. Nonetheless, a reduction in required sample size saves money and effort in the evaluation process.

Overall, whether to use the school or the teacher as the unit of randomization depends on the specific features of the intervention being tested and the potential for collaboration among instructional staff in the study schools. If the features of the intervention do not require school-level random assignment, and if control group contamination is not a major concern, then randomizing at the teacher level might be better. If, however, the intervention or school context is prone to cause contamination, then randomizing at the school level would be preferable.

#### *Design Issue 4: How Big Should the Sample Be?*

To determine the sample size, one needs a basis for deciding how much precision is needed. The precision of impact is assessed through power analysis and expressed in terms of the smallest program effect that could be detected with confidence, or minimum detectable effect (MDE).

From a programmatic perspective, the appropriate way to determine the MDE might be to decide whether the study can detect an effect that, judging from the performance of similar programs, is likely to be attainable. This “attainable effect” is

especially difficult to decide for PD interventions. Unlike other types of educational interventions, PD programs target teachers directly and only affect student outcomes through teachers indirectly. So it is not yet clear how much precision one needs when evaluating the impact of a PD intervention.<sup>16</sup> At a minimum, one should review the record of past studies with dosage levels similar to the dosage of the intervention one wishes to study (cf. Yoon et al., 2007).

Statistically, the MDE is defined as the smallest true program effect that has a certain probability of producing an impact estimate that is statistically significant at a given level (Bloom, 1995). This parameter, which is a multiple of the impact estimator's standard error (see the first formula in note 15), depends on the following factors:

- The type of test to be performed: A one-tailed  $t$  test is used for program impact in the predicted direction; a two-tailed  $t$  test can be used for any program effects.
- The level of statistical significance to which the result of this test will be compared ( $\alpha$ ).
- The desired statistical power ( $1-\beta$ ): the probability of detecting a true effect of a given size or larger.
- The degrees of freedom of the test, which depends on the number of clusters ( $J$ ) and the cluster size.
- The intraclass correlation: the proportion of the total population variance across clusters as opposed to within clusters.
- The explanatory power of potential covariates, such as student's prior achievement, teacher characteristics, and so on (Bloom, 2005).

On one hand, the first three factors have their conventional values and are relatively easy to pin down.<sup>17</sup> Taking into account the fourth factor, the MDE size declines in roughly inverse proportion to the square root of the number of clusters randomized. Interestingly, the size of the clusters randomized often makes far less difference to the precision of program impact estimators than does the number of clusters (for more details, see Schochet, 2005).

The last two factors, on the other hand, are hard to determine. These factors vary from outcome to outcome and from sample to sample. The conventional practice is to use similar measures from similar past studies as proxies for these factors in the precision calculation at the design phase of a study. However, it is often difficult to find good proxies, and judgments must be made when deciding values for these factors. This is especially true for evaluations of PD interventions that use teacher-level randomization because measures of teacher-level intraclass correlation and the explanatory power of teacher-level covariates are not commonly available in past studies.

The power analysis must, of course, be based on the analytic model to be employed. For a study with random assignment at the school level, a three-level hierarchical model can be used to reflect the nested nature of the data, with students nested within teachers nested within schools. For a study with random assignment at the teacher level—where teachers are assigned randomly in each school—a two-level estimation model can be used to reflect the nesting of students within teachers, with schools modeled as fixed effects. Alternatively, a three-level estimation model

can be used to account for the nesting data structure, with schools modeled as random effects. In such a model, the treatment would appear at the second level (i.e., the teacher level). Power calculations based on these two approaches to treating schools in teacher-level studies will differ slightly owing to the model specifications.

#### *Design Issue 5: What Should Be Measured, How, and When?*

A fifth methodological issue is deciding what to measure—and how and when to measure it. According to Supovitz (2001), three common weaknesses of PD effectiveness studies are poor alignment between the *pedagogy* that the PD trains teachers to use (e.g., inquiry-oriented instruction) and how the students are tested (e.g., multiple choice), poor alignment between the *content* of what is taught in the classroom and the content on which students are tested, and lack of sufficient *time lag* between the PD intervention and the measurement of PD impact. To overcome these weaknesses, several related measurement issues need to be considered during the design stage. In the discussion below, we emphasize (a) measuring key mediators, (b) determining the alignment of the outcome measures with the intervention, and (c) determining the timing of measurement of outcomes.

#### *Measurement of Mediating Variables*

An important design decision concerns how much of the study's resources to devote to the measurement of anticipated mediating factors, such as implementation levels achieved and proximal outcomes (e.g., teacher knowledge and practice), in addition to distal outcomes (student achievement). Although it is tempting in randomized field trials to focus only on the ultimate outcome of student achievement, including measures of proximal outcomes and other potential moderators and mediators can have significant payoff. Measurement of mediating variables is especially critical in making use of study results to draw conclusions about the theory of teacher change and the theory of instruction on which the PD intervention is based.

For example, if a study of PD does not find an overall impact on student achievement, and no teacher outcomes are measured, it is impossible to know where and to what extent the causal model broke down. It may be that the PD was effective in increasing teachers' knowledge or practices (so the theory of teacher change receives support), but these teacher changes did not result in higher student achievement (so the theory of instruction is not supported). Without a measure of the proximal outcomes of the PD, the model cannot be fully explored or understood.

Similarly, if dosage, or exposure to the PD, is not measured, it will be unclear whether (a) the PD was successfully delivered as intended but did not obtain the desired results or (b) the PD was not successfully implemented and thus the theory of teacher change broke down at the first link in the chain. For this reason, it is important to anticipate in advance what the most important features of the PD model are and to design measures to quantify them. The most critical factor will probably be dosage, but other aspects of the PD may be important to measure as well—in particular, the time allocated to the specific topics covered. The difficulty of measuring dosage is likely to vary with the form of

PD—hours of attendance at a training is very simple to measure, whereas participation in coaching activities in the school is more challenging.

### *Alignment of Outcome Measures*

PD interventions will vary in the specificity of the intended outcomes for teacher knowledge, teacher practice, and student achievement. In particular, as discussed earlier, some approaches to PD may be designed to improve teachers' skills in implementing a highly specified set of instructional practices; other PD may be designed to strengthen teachers' content knowledge and pedagogical content knowledge, with desired changes in practice less clearly articulated. Similarly, some PD interventions may be designed to produce changes in relatively narrowly defined aspects of student achievement (e.g., student understanding of a particular concept in mathematics), whereas other PD interventions may be designed to produce broad-gauge changes in achievement.

Clearly, the outcome measures must be chosen to reflect the intended outcomes of the PD. However, the desired degree of alignment can be difficult to establish, and it may be wise to choose multiple instruments that are more or less closely aligned with the specific focus of the PD to provide some information on the generalizability and robustness of the findings.

Two extremes can exemplify the need for balance between alignment and generalizability in outcome measure selection. Assume two studies have been conducted to determine whether PD on middle school science content and instruction leads to higher achievement. In both scenarios, the intervention was focused primarily on providing teachers with an understanding of Newton's laws for use in middle school science, and the researchers chose a teacher outcome measure that was created by the PD team to assess teachers' understanding of the content just learned. The first study chose the language and examples in the assessment directly from the PD training materials, similar to an end-of-chapter test found in many textbooks. In addition, the study created a similar student achievement measure that exclusively focused on the aspects of Newton's laws that the teachers had learned to address explicitly in their physical science instruction. Both measures included a range of easy to difficult items to ensure that ability at all levels was captured. On conducting the impact analyses, the study obtained an effect for student achievement of 1.20 standard deviations and concluded that PD in Newton's laws affects middle school physical science achievement.

The second study, examining the impact of the same intervention, relied instead on existing standardized assessments of teacher and student ability. These measures had been widely tested and validated and were commonly used in the education domain as accountability tests, so the results were policy relevant. In addition, the data collection was low burden for study participants because the tests were district administered and data were available from secondary sources. The tests measured a variety of middle school science outcomes but did not provide subscores on Newton's laws, which was the focus of the PD. On conducting an impact analysis, this group obtained null results.

The differences in results across the two studies may have little to do with the PD being studied and more to do with the

alignment of measures with the intervention. The first study chose a strongly aligned test; the second chose a weakly aligned test. If the intended outcomes of the PD are very narrowly and specifically designed, it may be sufficient to focus on a highly aligned set of measures; in most cases, though, it appears that a mix of measures varying in specificity will be required to permit an examination of the generalizability of the results.

Given the cost of measuring student achievement using well-aligned measures, we would argue that it is best to initially study the effect of a PD intervention on teacher knowledge only. The impact on knowledge can be measured at significantly less expense, both because of savings from eliminating the cost of measuring student achievement and because of savings from using teachers as the unit of assignment rather than whole schools. Moreover, to have an impact on student achievement of a detectable magnitude, the impact on teacher knowledge must be quite substantial.<sup>18</sup>

### *Timing of Outcome Measurement*

Some PD programs are designed to improve teacher practice and student achievement over the course of an entire year. But in principle, PD could focus on shorter increments of instruction. Regardless of the PD being studied, careful attention is needed to the timing of outcome measurement.

Timing is important because moving from providing PD to obtaining an impact on achievement involves traversing a number of causal links, and each of these may take time to unfold. How long do teachers need to think about what they have learned in order to put it effectively into practice? Once practices are put in place, are they sustained over time? And how long does improved instruction take to effect detectable increases in students' learning? That is, does PD have an impact on student outcomes during the year in which the PD is delivered, or is there a delay in impact? The results of the observational study described earlier suggested that content-focused PD received by middle school mathematics teachers may have an impact on student achievement in the year following participation in the PD but not during the year of participation (Harris & Sass, 2007).

The answer to these questions will likely vary by the intensity, specificity, and form of the PD received and by the alignment of the outcome measure to the PD. A focused weeklong institute on phonics with a lot of modeling and training on specific instructional practices may begin to affect teachers' practices as soon as the teachers return to the classroom. If the student test used to measure achievement is very sensitive to these practices, achievement gains might occur relatively quickly.

In most studies, however, these conditions are not likely to be met, suggesting that a realistic study will require multiple waves of data collection, with the timing determined by features of both the PD and the measures.

## **Conclusion**

Given the large public investment in PD, there is much to gain from research that addresses practical questions faced by those who design and adopt PD programs. The existing literature—as summarized by Yoon et al. (2007)—demonstrates that carefully constructed PD, delivered by its creators in conducive settings,

can have an effect on student achievement. But there is a need to demonstrate the effectiveness of such programs when delivered by others in a range of contexts. The literature also reveals an informal consensus about the features that PD programs should have in order to make them effective. But the evidence base for this consensus is weak, and there is a need to test specific features in order to address basic questions of policy and practice.

This article considers the benefits of experiments in addressing these research needs and offers a methodological resource. Experiments give researchers extensive control over what treatments to study, the contexts in which to study those treatments, the structure and size of the sample, and the timing and alignment of the measures to be used. Further discussions of these issues may be of use to those conducting and interpreting experiments involving PD.

## NOTES

We would like to thank our colleagues at American Institutes for Research and MDRC, especially Fran Stancavage, James Taylor, and Kirk Walters, whose thoughts over the course of different projects are incorporated in this manuscript. In addition, Willis Hawley and Marsha Silverberg provided helpful feedback on an earlier draft as discussants at the 2007 annual meeting of the Association for Public Policy Analysis and Management. We also received valuable comments on earlier drafts from Peter Youngs and from several anonymous reviewers. We take responsibility for any errors.

<sup>1</sup>Note that we focus on studies of professional development (PD) programs rather than on studies of teacher learning, broadly construed. There is a significant body of research on how teachers learn through nonprogrammatic mechanisms, such as informal, unplanned interaction with one's colleagues. See, for example, Bransford, Brown, and Cocking (1999).

<sup>2</sup>In this article, whenever we refer to effects on student achievement, the basis for the effect size is always one student-level standard deviation. In Yoon, Duncan, Lee, Scarloss, and Shapley (2007), the studies that reported effect sizes as a proportion of the student-level standard deviation were those by Cole (1992); Marek and Methven (1991); McCutchen et al. (2002); McGill-Franzen, Allington, Yokoi, and Brooks (1999); and Tienken (2003). The other studies reported apparently used the teacher-level standard deviation.

<sup>3</sup>Decker, Mayer, and Glazerman (2004) obtained an effect size of .26 in the analysis that focused on teachers with 3 or fewer years of experience. When all teachers were included in the analysis, the effect size was .15.

<sup>4</sup>Our intent in this paragraph is to show that evaluations of PD have found effect sizes that seem promising. We caution against interpreting these differences in effect sizes as evidence that PD is more effective than other interventions involving teachers. Effect sizes also depend, for instance, on the student outcome measures used and on the grade levels of the students. For more on effect sizes in education, see Konstantopoulos and Hedges (2008).

<sup>5</sup>In organizing the extensive literature on teacher PD, Borko (2004) posited three phases of research. Many existing studies had sought to prove the efficacy of specific programs delivered by a single trainer (Phase 1). Borko argued that the field needed more studies showing that such programs could be replicated more broadly (Phase 2). Finally, she argued that researchers should then test whether the replicated PD programs continued to be effective, in various settings and using various trainers, using either quasi-experimental or experimental research designs (Phase 3). This pipeline should arguably yield information that is highly relevant for policy and practice.

<sup>6</sup>In Kennedy's (1998) review, Group 1 PD programs were focused on the activities that prescribe a set of teaching behaviors that are expected to apply generically to all school subjects (e.g., Stevens & Slavin, 1995). Group 2 PD activities also prescribed a set of generic teaching behaviors but were prescribed for use with one particular school subject (e.g., Good, Grouws, & Ebmeier, 1983). Group 3 PD activities provided general guidance on both curriculum and pedagogy for teaching a particular subject, and their recommended practices were justified with references to knowledge about how students learn this subject (e.g., Wood & Sellers, 1996). Finally, Group 4 PD programs provided knowledge about how students learn particular subject matter but did not provide specific guidance on the practices that should be used to teach that subject (e.g., Carpenter, Fennema, Peterson, Chiang, & Loef, 1989).

<sup>7</sup>Collectively, the studies that Yoon et al. (2007) reviewed suggested that the duration or "dosage" of PD may be related to impact. The average number of contact hours was about 48 hours across the nine studies, and the three studies that provided the least intensive PD (ranging from 5 to 14 hours over the span of 2 to 3½ months) produced no statistically significant effect. The remaining six studies provided more intensive PD, with contact hours ranging from 30 to more than 100. With the exception of one study with 83 contact hours, all the studies with 30 or more hours of PD resulted in significant effects on student achievement.

<sup>8</sup>It is worth noting that such techniques are much more commonly used in observational studies than in quasi-experiments. The techniques require relatively large samples, which are more typical in observational studies.

<sup>9</sup>We have not identified all the relevant techniques, nor have we listed all the limitations of those identified here. For a more thorough discussion of these techniques and their limitations, see Appendix 5.1 in Shadish, Cook, and Campbell (2002).

<sup>10</sup>Sloan (1993) reported statistically significant effects on student achievement outcomes measures. In a subsequent analysis of Sloan's findings, Yoon et al. (2007) adjusted for clustering and for tests on multiple outcomes and found that impacts on student achievement were large enough to be substantively important but were not statistically significant.

<sup>11</sup>Note that three of the authors of this article work on the Study of the Impact of Two Professional Development Interventions in Early Reading.

<sup>12</sup>In addition, in a three-arm study the two treatments might be designed to focus on different mediating variables. For example, one treatment might be designed to improve teachers' content knowledge, and the other might be designed to focus more directly on changing instructional practice. This variation might induce variation in the mediators, making it possible to disentangle the effects of the mediators on outcomes, using instrumental variable techniques.

<sup>13</sup>We have supposed that the PD does not involve the introduction of curricular materials that displace existing materials. The requirements placed on the context will differ if new materials are to be introduced along with the PD; in that case, one would need to consider the contrast between the curriculum in use and the new materials to be adopted.

<sup>14</sup>In a school-level design, turnover would not ordinarily lead to selection bias even if there were differential turnover between treatment and control schools, because all teachers in the school would be included in the analysis. However, the estimated impact would consist of a combination of the impact of the treatment on achievement for teachers who stay and the impact on turnover.

<sup>15</sup>The standard error of the impact estimator for cluster randomization (when no covariates other than the treatment status is included)

is  $SE_{CL} = \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\tau^2}{J} + \frac{\sigma^2}{nJ}}$ , whereas the standard error of the impact

estimator for individual randomization program is  $SE_{IN} = \sqrt{\frac{1}{P(1-P)} \left( \frac{\tau^2}{nJ} + \frac{\sigma^2}{n} \right)}$ .

Here,  $P$  is the percentage of clusters (or individuals) that are randomly assigned to the treatment group;  $J$  is the number of clusters;  $n$  is the number of individuals within a cluster;  $\tau^2$  is the cross-cluster variance, and  $\sigma^2$  is the within-cluster variance (Bloom, 2005). The proportion of the total population variance across clusters as opposed to within clusters ( $\frac{\tau^2}{\tau^2 + \sigma^2}$ ) is usually called an intraclass correlation (Fisher, 1925).

<sup>16</sup>Because a study of PD is likely to focus on measures at both the teacher and student levels, at least two different minimum detectable effect sizes (MDES) should be examined—one for teacher outcomes and one for student outcomes. More work is needed on the relationship between the desired MDES at the teacher and student levels—that is, the minimum effect required at the teacher level to produce the minimum effect desired at the student level.

<sup>17</sup>Other than in efficacy studies, researchers usually employ a two-tailed test with a statistical power of 80% and significance level of .05.

<sup>18</sup>If the percentage of the variance in student achievement that lies within teachers is large, then a large effect must be obtained at the teacher level (as measured in terms of the between-teacher standard deviation) to produce a modest effect at the student level (as measured in terms of the between-student standard deviation).

## REFERENCES

- Birman, B., Le Floch, K. C., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., et al. (2007). *State and local implementation of the No Child Left Behind Act: Vol. 2. Teacher quality under NCLB: Interim report*. Washington, DC: U.S. Department of Education; Office of Planning, Evaluation and Policy Development; Policy and Program Studies Service.
- Blank, R. K., de las Alas, N., & Smith, C. (2008). *Does teacher professional development have effects on teaching and learning? Analysis of evaluation findings from programs for mathematics and science teachers in 14 states*. Washington, DC: Council of Chief State School Officers.
- Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, 19(5), 547–556.
- Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). New York: Russell Sage Foundation.
- Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, 33(8), 3–15.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.
- 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(4), 499–531.
- Choy S. P., Chen, X., & Bugarin, R. (2006). *Teacher professional development in 1999–2000: What teachers, principals, and district staff report*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Clewell, B. C., Campbell, P. B., & Perlman, L. (2004). *Review of evaluation studies of mathematics and science curricula and professional development models*. Washington, DC: Urban Institute.
- Cole, D. C. (1992). The effects of a one-year staff development program on the achievement test scores of fourth-grade students. *Dissertation Abstracts International*, 53(06), 1792A. (UMI No. 9232258)
- Decker, P., Mayer, D., & Glazerman, S. (2004). *The effects of Teach for America on students: Findings from a national evaluation*. Princeton, NJ: Mathematica Policy Research.
- Desimone, L., Porter, A. C., Garet, M., Yoon, K. S., & Birman, B. (2002). Does professional development change teachers' instruction? Results from a three-year study. *Educational Evaluation and Policy Analysis*, 24(2), 81–112.
- Donner, A., & Klar, N. (2000). *Design and analysis of group randomization trials in health research*. London: Arnold.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.
- Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Education Research Journal*, 38(4), 915–945.
- Good, T. L., Grouws, D. A., & Ebmeier, H. (1983). *Active mathematics teaching*. New York: Longman.
- Guskey, T. R. (2003). What makes professional development effective? *Phi Delta Kappan*, 84(10), 748–750.
- Harris, D. N., & Sass, T. R. (2007). *Teacher training, teacher quality, and student achievement*. Unpublished manuscript, University of Wisconsin–Madison.
- Hawley, W. D., & Valli, L. (1998). The essentials of effective professional development: A new consensus. In L. S. Darling Hammond & G. Sykes (Eds.), *The heart of the matter: Teaching as a learning profession* (pp. 86–124). San Francisco: Jossey-Bass.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Joyce, B., & Showers, B. (2002). *Student achievement through staff development* (3rd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
- Kellam, S. G., & Langevin, D. J. (2003). A framework for understanding “evidence” in prevention research and programs. *Prevention Science*, 4(3), 137–153.
- Kennedy, M. (1998). *Form and substance of inservice teacher education* (Research Monograph No. 13). Madison: University of Wisconsin–Madison, National Institute for Science Education.
- Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reform? *Teachers College Record*, 110(8). Retrieved July 8, 2008, from <http://www.tcrecord.org>
- Little, J. W. (1993). Teachers' professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis*, 15(2), 129–151.
- Loucks-Horsley, S., Hewson, P. W., Love, N., & Stiles, K. E. (1998). *Designing professional development for teachers of science and mathematics*. Thousand Oaks, CA: Corwin Press.
- Marek, E. A., & Methven, S. B. (1991). Effects of the learning cycle upon student and classroom teacher performance. *Journal of Research in Science Teaching*, 28(1), 41–53.
- McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, S. N., Cox, S., Potter, N. S., et al. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities*, 35(1), 69–86.
- McGill-Franzen, A., Allington, R. L., Yokoi, L., & Brooks, G. (1999). Putting books in the classroom seems necessary but not sufficient. *Journal of Reading Research*, 93(2), 67–74.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.
- Murray, D. M., Hannan, P. J., Jacobs, D. R., McGovern, P. J., Schmid, L., Baker, W. L., et al. (1994). Assessing intervention effects in the Minnesota heart health program. *American Journal of Epidemiology*, 139(1), 91–103.

- National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future*. New York: Author.
- Saxe, G. B., Gearhart, M., & Nasir, N. S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. *Journal of Mathematics Teacher Education*, 4(1), 55–79.
- Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs*. Princeton, NJ: Mathematica Policy Research.
- Shadish, W. R., Cook, T. D., & Campbell, T. D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Showers, B., Joyce, B., & Bennett, B. (1987). Synthesis of research on staff development: A framework for future study and a state-of-the-art analysis. *Educational Leadership*, 45(3), 77–87.
- Sloan, H. A. (1993). Direct instruction in fourth and fifth grade classrooms. *Dissertation Abstracts International*, 54(08), 2837A. (UMI No. 9334424)
- Society for Prevention Research. (2004). *Standards of evidence: Criteria for efficacy, effectiveness, and dissemination*. Fairfax, VA: Author.
- Stevens, R. J., & Slavin, R. E. (1995). The cooperative elementary school: Effects on student achievement, attitudes, and social relations. *American Educational Research Journal*, 32, 321–351.
- Supovitz, J. A. (2001). Translating teaching practice into improved student achievement. In S. Fuhrman (Ed.), *National Society for the Study of Education yearbook*. Chicago: University of Chicago Press.
- Supovitz, J. A., Mayer, D., & Kahle, J. B. (2000). The longitudinal impact of inquiry-based professional development on teaching practice. *Educational Policy*, 14(3), 331–356.
- Tienken, C. H. (2003). The effect of staff development in the use of scoring rubrics and reflective questioning strategies on fourth-grade students' narrative writing performance. *Dissertation Abstracts International*, 64(02), 388A. (UMI No. 3081032)
- Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. *Review of Research in Education*, 24, 173–209.
- Wood, T., & Sellers, P. (1996). Assessment of a problem-centered mathematics program: Third grade. *Journal for Research in Mathematics Education*, 27, 337–353.
- 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.
- Yoon, K. S., Garet, M., Birman, B., Jacobson, R. (2006). *Examining the effects of mathematics and science professional development on teachers' instructional practice: Using professional development activity log*. Washington, DC: Council of Chief State School Officers.

## AUTHORS

**ANDREW J. WAYNE** is a senior research analyst at the American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007; [awayne@air.org](mailto:awayne@air.org). His research focuses on policies and programs that affect teachers.

**KWANG SUK YOON** is a principal research analyst at the American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007; [kyoon@air.org](mailto:kyoon@air.org). His research interests include the quality and effects of teacher professional development and program evaluations.

**PEI ZHU** is a senior research associate at MDRC's K–12 education policy area, 19th Floor, 16 East 34 Street, New York, NY 10016-4326; [pei.zhu@mdrc.org](mailto:pei.zhu@mdrc.org). Her research focuses on experimental and quasi-experimental impact analyses, evaluation design, and related methodological issues.

**STEPHANIE CRONEN** is a principal research analyst at the American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007; [scronen@air.org](mailto:scronen@air.org). Her research focuses on the effects of a range of educational interventions.

**MICHAEL S. GARET** is chief scientist at the American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007; [mgaret@air.org](mailto:mgaret@air.org). His research focuses on research methods, schools as organizations, and teacher professional development.

Manuscript received July 17, 2008

Revision received September 15, 2008

Accepted September 28, 2008