Review of Educational Research
Month 201X, Vol. XX, No. X, pp. 1–36
DOI: 10.3102/0034654315626800

How Does Professional Development Improve Teaching?

Mary M. Kennedy
Michigan State University

Professional development programs are based on different theories of how students learn and different theories of how teachers learn. Reviewers often sort programs according to design features such as program duration, intensity, or the use of specific techniques such as coaches or online lessons, but these categories do not illuminate the programs' underlying purpose or premises about teaching and teacher learning. This review sorts programs according to their underlying theories of action, which include (a) a main idea that teachers should learn and (b) a strategy for helping teachers enact that idea within their own ongoing systems of practice. Using rigorous research design standards, the review identifies 28 studies. Because studies differ in multiple ways, the review presents program effects graphically rather than statistically. Visual patterns suggest that many popular design features are not associated with program effectiveness. Furthermore, different main ideas are not differentially effective. However, the pedagogies used to facilitate enactment differ in their effectiveness. Finally, the review addresses the question of research design for studies of professional development and suggests that some widely favored research designs might adversely affect study outcomes.

Keywords: professional development, teacher learning, research design, teacher quality, educational improvement, teaching practice

The idea that professional development (PD) can foster improvements in teaching is widely accepted. PD is required by virtually every teaching contract in the country, and teachers participate in PD every year. Foundations and federal agencies spend large sums on the design and implementation of PD programs. Yet despite this widespread agreement about its importance, there is little consensus about how PD works, that is, about what happens in PD, how it fosters teacher learning, and how it is expected to alter teaching practice. The actual form and substance of PD programs is tremendously various, raising questions about why something so various is uniformly assumed to be a good thing. This article reviews research on PD programs with an eye toward learning more about how different approaches to PD actually foster learning. Taking up this project raises a host of

related questions about what teachers need to learn, what kind of PD activities foster learning, and about how learning in one context, such as a PD workshop, affects behavior in another, such as a teachers' own classroom.

Because PD programs are so various, reviewers generally try to sort them according to similarities and differences in their visible features. Some reviewers focus on differences in content. For instance, Sher and O'Reilly (2009) compared programs that focused on subject matter with those that focused on pedagogy, and Kennedy (1998) identified four substantive foci: generic teaching practices, subject-specific teaching practices, curriculum and pedagogy, and how students learn. Other reviewers focus on features of program design. For instance, Blank and de las Alas (2009) found that more effective programs have features such as follow-up steps in teachers' schools, active learning methods, collective participation, and substantive attention to how students learn specific content. Timperley, Wilson, Barrar, and Fung (2007) categorized 84 design features of PD programs including the content and process of the PD programs, the characteristics of the teachers' schools, and the social context of their work.

These lists of salient design features and operational mechanisms help simplify the complex array of programs, but they have also been criticized. One group of reviewers (Sztajn, Campbell, & Yoon, 2011) criticized these lists for shifting our attention away from relevant conceptual and theoretical frameworks, and another (Opfer & Pedder, 2011) criticized them for being based in process–product logic, meaning that programs are defined by visible processes or features rather than by the functions these processes serve.

One reason we rely on such lists is that there is no single, overarching theory of teaching or of teacher learning. Teachers are characterized as managers, actors, mediators, role models, salesmen, and so forth. With different conceptions of teaching come different conceptions of how PD can improve teaching. But we cannot learn from this body of research unless we find a way to distinguish among these different conceptions of what teachers are actually doing and how we can help them improve. Rather than considering the array of specific design features PD programs rely on, this review seeks to define programs according to their underlying theories of action, where a theory of action includes two important parts. First, it identifies a central problem of practice that it aims to inform, and second, it devises a pedagogy that will help teachers enact new ideas, translating them into the context of their own practice.

What Problems of Practice Do Programs Aim to Inform?

Questions about what teachers need to know are typically prefaced by stipulations about what teachers actually do. Ever since Charters and Waples (1929) first tried to define teachers' practices, researchers have struggled to find a common language that could define teachers' work and delineate the knowledge needed to guide that work. A large body of research in the 1970s, commonly dubbed "process–product" research, sought to distinguish relatively more and less successful practices. This was followed in the 1980s by a surge of close-up studies of teacher thinking, nicely summarized by Clark and Peterson (1986), designed to help us understand how teachers reasoned about and made decisions about their practices. From these studies, we learned that teachers are continuously balancing among

multiple and conflicting goals and ideals, some self-imposed and some imposed from outside. For example, as a society we expect teachers to treat all children equally, yet respond to each child's unique needs; to be strict yet forgiving; and to be intellectually demanding yet leave no child behind.

Along with these widespread public ideals, teachers also commit to a variety of personal ideals and visions of their ideal selves. They may strive to be fair, to have a sense of humor, to maintain a quiet voice, to provide at least one encouraging word to each student every day, or to make sure they give time to their most needy students. Local districts also add demands, requiring teachers to add a curriculum unit on the history of their town, or asking them to enforce a new hats-off policy. Students, too, can place demands on teachers. A widely recognized phenomenon in classrooms is students who bargain with teachers to reduce the intellectual challenge of their assigned tasks (Metz, 1993; Sedlak, Wheeler, Pullin, & Cusick, 1986).

Thus, from the teachers' perspective, the education system is "noisy": Teachers are surrounded by multiple and conflicting messages about what is most important to do. Furthermore, if they focus too much on any one of these important ideals, they may compromise their effectiveness with another. For instance, methods used to contain student behavior can become so heavy-handed that they reduce students' motivation to participate in the lesson (McNeil, 1985), or methods used to make lessons more interesting or entertaining can inadvertently distort the content itself (Doyle, 1986). In his classic study of American high schools, Cusick (1983) argued that maintaining higher academic standards was only marginally important in high schools and that their foremost goal was to contain the behaviors of students who did not want to be there. Similarly, Kennedy (2005) found that, although elementary teachers wanted their students to be actively engaged in learning, there was a limit to how much engagement they would tolerate, for too much engagement led to too much noise and classroom frenzy.

Because teachers' work is inherently multifaceted and driven by a wide variety of conflicting ideals and the ideas, any review of PD should attend to the particular ideas programs offer to teachers and the particular aspect of practice they hope to improve. Later, in the "Method" section of this article, I identify four persistent challenges of practice that I used to classify the goals of PD programs.

What Pedagogy Do Programs Use to Facilitate Enactment of Their Ideas?

The second important feature of a PD theory of action has to do with how it helps teachers translate new ideas into their own systems of practice. This is important because PD programs typically meet with teachers *outside* of their classrooms to talk about teaching, yet they expect their words to alter teachers' behaviors *inside* the classroom. They are at risk for what Kennedy (1999) called the *problem of enactment*, a phenomenon in which teachers can learn and espouse one idea, yet continue enacting a different idea, out of habit, without even noticing the contradiction. Furthermore, because PD providers work with practicing teachers, they are by definition not merely offering a *new* idea but rather a *different* idea from the one that has guided teachers in the past. Teachers participating in PD have already developed their practice and they have already found ways to

balance among their many competing challenges and ideals. They are likely to have formed habitual responses to students jumping out of their seats, to favor certain methods of portraying particular curriculum content, to favor certain seating arrangements, bulletin board displays, and so forth. Thus, any new idea offered by PD requires not merely adoption but also *abandonment* of a prior approach. Later, in the "Method" section, I describe four methods of facilitating enactment that I found in the studies reviewed here.

So this review characterizes PD theories of action by their main ideas and by the way they help teachers enact those ideas. In addition, there is also an important feature of PD research designs that is addressed in this review. A central problem in PD research is ensuring that teachers in program and comparison groups are actually comparable. When random assignment is not possible, a common method for identifying a comparison group for PD studies is to select teachers whose teaching assignments match those of program teachers. Comparison teachers teach the same grade levels, same subjects, and/or same types of students. But this strategy fails to match teachers on their motivation to learn. Typically, participating teachers *chose* to participate in PD, whereas comparison teachers with similar teaching assignments did not. Hundreds of studies have been done using comparison groups whose only similarity to the program teachers is their teaching assignment. This review does not include such studies, on the ground that researchers have not ensured comparable motivation to learn, and this is the most important form of comparability when a study is about learning.

On the other side, there are also situations in which teachers are required to participate in PD. When studies mandate teachers' assignment to treatment and comparison groups, they ensure that their two groups are comparable, but they do so by ensuring that *neither group* is motivated to learn. Because the effects of any PD program depends heavily on teachers' motivation to learn and to change their practice, studies using mandatory assignments may not have much effect on learning. This is not to say that teachers will actively resist but rather that they will forget about the program when they return to their classrooms. This review aims to learn more about how PD programs and studies address these challenges and whether their solutions make a difference to student learning. The next section describes methods for searching the literature and analyzing studies.

Method

This review examines experimental studies of PD carried out in K-12 general education within the United States and published since 1975. The restriction to the United States acknowledges that teaching is inherently cultural and varies across nations (Hiebert et al., 2003; Roth et al., 2006; Stigler & Perry, 1988) and that the United States is somewhat unique in its lack of a national curriculum. The 1975 time limit reflects the history of this field: Few, if any, experimental studies of PD were carried out prior to that time. The review focuses on K-12 teachers teaching core academic subjects (language arts, mathematics, the sciences, and the social sciences). It does not include studies conducted in special subjects such as art, music, physical education, agriculture, and so on, where pedagogical demands can be quite different.

Search for Studies

The topic of PD is very popular. There could be thousands of articles written about it almost every year, but the vast bulk of these articles do not present experimental evidence. Some simply describe programs. Some include teacher testimony. Others rely on classroom observations and still others measure student learning. Some include a comparison group; others do not. Many studies combine PD with other supports such as new curriculum materials or new technology. Because of the large volume of work that is relevant but not eligible for this review, I did not undertake a database search. Such searches are far too labor-intensive relative to their yield. Instead, I first sought candidates from other reviews of PD, including Blank, de las Alas, and Smith (2008); Borko (2004); Hirsh and Hord (2008); Kennedy (1998); Lampert (1988); Loucks-Horsley and Matsumoto (1999); Mitchell and Cubey (2003); Opfer and Pedder (2011); Sher and O'Reilly (2009); Timperley et al. (2007); Wei, Darling-Hammond, and Adamson (2010); Wideen, Mayer-Smith, and Moon (1998); and Yoon, Duncan, Lee, Scarloss, and Spapley (2007).

These reviews appeared to adequately capture research prior to 2000, when formal studies of PD were relatively rare. To ensure coverage after 2000, I hand-searched journals that cover teacher education–related topics. I examined all issues published from 2000 to 2014 from the following journals: *American Educational Research Journal*, *Journal of Educational Psychology*, *Journal of Teacher Education*, *Teaching and Teacher Education*, *Teachers College Record*, *Peabody Journal of Education*, *American Journal of Education*, *The Elementary School Journal*, *Journal of Research in Mathematics Education*, *Journal of Research in Science Teaching*, *School Science and Mathematics*, and *Reading Research Quarterly*. Finally, I examined the reference lists in the articles that I found. Even after eliminating studies using the criteria outlined below, this approach yielded 28 studies. This is far more than two other best-evidence syntheses, each of which yielded fewer than 10.

Criteria for Study Selection

To ensure that the studies offered valid, and comparable, inferences about PD, I required them to meet five criteria. These are described below.

The Study Is About PD Only

One thing that makes a review of PD difficult is that researchers use PD to study many things other than the PD itself. That is, a researcher may be interested in a new curriculum, a new teaching technique, or new classroom tool, but must provide PD to enable teachers to use their innovation. To learn about the benefit of the PD, itself, researchers need to make sure that participation in PD is the *only* difference between groups. For, as Slavin, Lake, Hanley, and Thurston (2014) pointed out, if teachers in two different PD programs are also teaching two different curricula, then the effects of the PD are confounded with the effects of the curriculum, so that differences in student achievement may actually reflect the different curricula, not the PD per se.

Using this criterion, I excluded, for instance, Newman et al.'s (2012) study of the new Alabama Math, Science, and Technology Initiative as well as studies in which PD was used to test the merits of instructional tools such as *Simcalc* (Roschelle et al., 2010) or the merits of a computerized student assessment system (e.g., Fuchs, 1991). I did allow curriculum-oriented PD studies when a comparison group taught the same new curriculum without the benefit of the PD (e.g., Borman, Gamoran, & Bowdon, 2008; Penuel, Gallagher, & Moorthy, 2011; Saxe, Gearhart, & Nasir, 2001).

The Study Includes Evidence of Student Achievement

I rejected studies that did not provide evidence of student achievement, even if they provided other forms of evidence such as teacher testimony, interviews, surveys, or classroom observations. There are two reasons for this. First, the ultimate goal of PD is to improve student learning. Second, measures of student learning are relatively more similar across studies than are classroom observations or teacher interviews, which are often tailored to capture activities uniquely relevant to a PD program. However, measures of student achievement can also differ. Conventional standardized tests and state assessments tend to cover a broader array of content and may be less sensitive to specific program purposes. Consequently, some researchers develop measures that are intentionally designed to capture unique program effects. In the language of Ruiz-Primo, Shavelson, Hamilton, and Klein (2002), these tailored instruments are more *proximal* to the program, whereas district or state assessments are more *distal*. This review includes both types of measures, but, as others have done (e.g., Roschelle et al., 2010), it presents conventional measures such as state assessments or standardized test scores as "M1" outcomes and those created specifically for the PD program as "M2" outcomes. M2 measures are more closely aligned with the program goals and might therefore be expected to demonstrate greater program impact.

The Study Design Controls for Motivation to Learn

I included studies that used mandatory assignments to groups, even though these might reduce motivation to learn, on the ground that the groups are nonetheless comparable in their motivation. The remainder of the studies used a variety of approaches to ensure teacher motivation even though teachers would be participating in different programs. Some invited teachers to participate and told them in advance that this was a study and they might not be assigned to the program they preferred. Some randomly assign volunteers to cohorts, so that everyone eventually gets the program but some get it a year later. One study (Glazerman et al., 2010) mandated *school* assignments and then *offered* the program to teachers in treatment schools. This procedure could have biased the study outcome, but the authors measured program impact on all teachers in the schools, whether or not they availed themselves of the service. Yet another approach (used by Roth et al., 2011) was to offer two different programs and let teachers to self-select between them. This approach would likely be rejected by conventional research standards; however, it does assure that both groups are equal in their motivation to learn *that particular program*. All these variations are included here.

A Minimum Study Duration of 1 Year

An important and underemphasized question in research on PD is whether PD produces *enduring* changes in practice rather than temporary compliance. Several studies with appropriate comparison groups and outcome measures were rejected because they followed teachers and their students for only a few weeks or months (e.g., Siegle & McCoach, 2007).

However, in many cases, the PD itself also extended throughout an entire school year, so that even year-long measures of student learning were still coterminous with the PD itself. Ideally, PD research should follow teachers for at least a full year *after the completion of the PD itself*, to discover the extent to which teachers are able to sustain the new practice after the PD support is gone. However, such a requirement would reduce the population of eligible studies to just a handful.

Researchers Follow Teachers, Rather Than Students, Over Time

Long-term studies present unique problems to educational researchers because teachers and students are reallocated each year and there is some confusion about whether researchers should follow the students or the teachers during ensuing school years. In one study (Heller, Dahler, Wong, Shinohara, & Miratrix, 2012), the researchers assessed long-term impact by following the *students* over time to see whether they still recalled the content they had originally learned from the program teachers. This study design tests whether teachers had an enduring impact on their *students*, but not whether the program had an enduring impact on the *teachers*. In another study (Campbell & Rowan, 1995), researchers compared program impact on *students* who had been in the program for 1, 2, or 3 years, rather than comparing *teachers*' performance when teachers had participated for 1, 2, or 3 years. The question of interest is in a PD study should be whether teachers themselves can sustain their improved practices in the ensuing years.

How These Rules Differ From What Works Clearinghouse

The criteria described above are intended to yield strong-inference studies of PD but they differ in important ways from those used by the What Works Clearinghouse (WWC). For instance, the WWC's 2007 review of PD literature (Yoon et al., 2007) included only nine studies, four of which would have been rejected using the criteria listed above. One of theirs (Duffy et al., 1986) was rejected from this review because the study duration was only 6 months. Another (McCutchen, Abbott, Green, & Beretvas, 2002) was rejected because even though the authors claimed to randomly assign teachers to treatment and comparison conditions, they also said that they gave certain schools preferential access to the treatment condition, a decision that could bias the study outcome. Two others (Marek & Methven, 1991; Tienken & Achilles, 2003) were rejected because they selected a comparison group by matching comparison teachers to the treatment teachers, a practice that fails to equate groups on their motivation to learn. On the other side, this review includes studies that might have been rejected by the WWC because of their unconventional assignment processes.

Estimating Individual Program Effects

The first step in analysis was to compute effect sizes for each student-learning outcome in each study. Computations depended on the original approach to data analysis. Most of the older studies used analysis of variance or covariance, but some used student-level data; others used classroom-level data. More recent studies typically relied on hierarchical linear models. These different approaches yield very different error terms, and hence different estimates of effect sizes. Below are my methods of establishing effect sizes from the different kinds of data.

Student-Level Data

The first group of studies used student-level data, analyzed with analysis of variance (ANOVA), analysis of covariance (ANCOVA), or *t* tests. Following WWC (2013), I used Hedge's *g* to calculate effect sizes, where

$$\overline{ES} = \frac{\overline{X_{G1}} - \overline{X_{G2}}}{s_{\text{pooled}}}$$

and

$$s_{\text{pooled}} = \sqrt{\frac{s_1^2\left(n_1 - 1\right) + s_2^2\left(n_2 - 1\right)}{n_1 + n_2 - 2}}.$$

This estimate tends to be upwardly biased for small samples, however, and many of these older studies used relatively small samples. There is a correction for small samples, which multiplies effect size estimates by

$$\left[1 - 3/\left(4N - 9\right)\right].$$

This correction is not sufficient, however, when the researcher is interested in teachers, for even though these studies had small samples of teachers, they had much larger samples of students, so that corrections for small samples have negligible effects. Yet some correction is needed because students within each classroom are not independent units and the researchers have not accounted for dependencies within each classroom. To correct for these likely upwardly biased estimates, I based Hedge's correction on the number of teachers rather than the number of students. This correction is conceptually appropriate, even if not statistically customary.

Classroom-Level Data

The second group of studies also relied heavily on ANOVA or ANCOVA, but used classrooms as their units of analysis. This analytic approach is conceptually correct, because the unit being "treated" is the teacher; however, variations among classrooms are much smaller than variations among individual students and so effect sizes tended to be very large, often more than full standard deviation above their comparison groups. None of these authors provided enough data to enable an

estimate of student-level standard deviations. I therefore devised an ad hoc approximation by *dividing each effect size by four*. My reasoning was as follows. First, variance decomposition studies (see, e.g., Rowan, Correnti, & Miller, 2002) suggest that variance among classrooms is roughly 15% to 20% of total variance, whereas variance among students is roughly 60% to 70%. Roughly, then, classroom variance is about a quarter the size of student variance. Second, within the sample of studies included in this review, the average effect size based on classroom units was .8 whereas the average effect based on student units was .19, roughly one fourth of the classroom effect size. After correcting all estimates for small samples and then dividing each class-level effect by 4, the average class-level effect was .17, and the average student level effect was .18.

Hierarchical Data

The third and largest group of studies used hierarchical models and these also varied considerably. Sometimes, individual teachers were allocated to treatments, sometimes, whole schools were, and sometimes, learning communities within schools were. The analytic approaches used for these studies were more complex than those in the first two groups and many were idiosyncratic to the studies. In most cases, I accepted the effect size estimates that authors offered. However, when studies yielded effects that appeared to be outliers, I contacted the authors to clarify the basis for their estimates and revised them if needed.

Within-Study Synthesis

Within each study, I calculated effects for every subgroup of students and/or every outcome measure and then averaged these to yield a single average outcome for each study. However, I retained distinctions between M1 and M2 measures so that readers can see differences in effect across these types of outcomes, and I retained distinctions across program years so that readers can see the differences between immediate and delayed program effects.

Comparisons Across Programs

The programs reviewed here are too various to benefit from a formal meta-analysis. Aggregating by any one dimension introduces important confounding on another dimension and renders interpretations difficult. I rely instead on visual displays that enable readers to see the unique effects of each program and to compare them individually with one another.

Characterizing PD Programs

To articulate similarities and differences among PD programs, I sought a classification system that would sort programs according to the two central aspects of their theories of action. Specifically, I categorized the content of the ideas they provided and their method for facilitating enactment of those ideas.

Program Content

I sorted program content according to the type of teaching problems they addressed, using a framework devised by Kennedy (2016). These are challenges that virtually every teacher must address, as they are inherent in the work of

teaching, and most PD programs are designed to address one of them. The first of these, and the one of most interest to observers, is portraying curriculum content in a way that enables naïve minds to comprehend it. If students could learn curriculum content simply by reading textbooks there would be no need for teachers. But they need help making sense of new content. Thus, we see teachers provide demonstrations, pictures, movies, hypothetical problems, walked-through examples, and so forth, and we see them devising learning activities for students to engage in on their own. For many observers of teaching, these activities are the essence of teaching, and we cannot say that someone is teaching if they are not portraying content to students.

The second persistent challenge is to contain student behavior. Students are energetic, excitable, boisterous, and easily distracted by one another. Teachers need to contain their behavior in part as a matter of safety but also to ensure that students are not distracting each other, or the teacher, from the lesson. The need to contain student behavior is acknowledged in virtually every school district performance assessment, where one or more items have to do with classroom management.

The third persistent challenge facing teachers is to enlist student participation. Teachers face a captive audience, and sometimes a resistant audience. The problem here is that *school attendance is compulsory but learning is not*. And as learning theorists remind us (see, e.g., Bransford, Brown, & Cocking, 1999) learning cannot occur without active intellectual engagement. This situation creates a predicament for teachers, for, as Cohen (2011) argued, teachers cannot claim to be teaching if students are not learning, yet students will not learn unless they themselves choose to actively participate in their lessons.

Finally, teachers must find ways to expose their students' thinking. This fourth persistent challenge may seem less obvious to observers than the first three, but without knowing what students understand at any given moment, teachers cannot know what to repeat, what to elaborate, or when to move on. Thus, we see teachers asking students to solve problems, share their findings, respond to one another's ideas, read aloud, show their work, turn in assigned projects for review, and so forth. Most school districts try to help teachers address this problem by providing them with formal assessment data, but the most useful knowledge for teachers is the knowledge they have *in the moment*, for this knowledge can guide their actions in the moment.

Thousands of hours have been spent arguing about the relative importance of these persistent teaching challenges, but all are fundamental to teaching, and teaching success depends on all four. Teachers cannot be said to be teaching unless students are learning and students cannot learn unless teachers portray content in a way that is comprehensible to naïve thinkers, enlist student participation in the lessons, contain distracting behavior among students, and expose student thinking so that they can adjust their lessons accordingly.

Notice, too, that these problems must be addressed simultaneously and continuously. Each new student, each new group of students, and each new topic to be taught requires teachers to think anew about how they will contain student behavior in this new situation, how they will enlist participation in this new situation, how they will portray curriculum content in this new situation, and how

they will expose their students' thinking in this new situation. These four persistent challenges provide a useful framework for characterizing PD programs. They are broad enough to contain a wide variety of program ideas, yet different enough to allow meaningful comparisons. Most program content fell easily into one or another of these categories.

Facilitating Enactment

The second part of a PD theory of action consists of a strategy for helping teachers enact new ideas within their own ongoing systems of practice. We have always confronted philosophical questions about how knowledge influences behavior, of course, but the question is further complicated in the case of teaching because teachers have already developed systems of practice that they believe optimally resolve the various challenges they face. For teachers, enacting a new idea is not a matter of simple adoption but rather a matter of figuring out whether, when, and how to incorporate that new idea into an ongoing system of practice which is *already satisfactory*, and may also be largely habitual. I identified four methods used by PD programs to facilitate enactment of their ideas.

The oldest and still most widespread method is prescription. Here, PD programs explicitly describe or demonstrate what they believe is the best way for teachers to address a particular teaching problem. From the PD provider's perspective, prescriptions reduce the amount of individual discretion or judgement that is needed, thus ensuring that teachers do things exactly as the provider intends. Furthermore, teachers are accustomed to receiving prescriptions. They routinely receive prescriptive guidance in the form of new laws, new school organizational mechanisms, new curricula, schedules, discipline policies, assessment systems, and record-keeping systems. Many popular and widely used commercial PD programs are also heavily prescriptive. Prescriptions are typically presented as universal, reducing the amount of flexibility or personal judgement teachers will need to enact the idea. But prescriptions can backfire if they address only one of the challenges teachers face, for in so doing they may exacerbate others.

Another way of facilitating enactment is through strategies. Strategies are distinctive in that they define their goals. Typically, PD programs convey a specific goal that teachers should strive for and then provide a collection of illustrative practices that will achieve that goal. The practices themselves can be just as procedurally detailed as prescriptions, but they differ in that they are accompanied by a rationale that helps teachers understand when and why they should implement these strategies. The challenge for PD is to make sure teachers understand the ultimate goal well enough that they can decide independently when they will use each strategy.

The third way to facilitate enactment is through insight. By definition, insights arise from self-generated "aha!" moments, but programs can foster new insights by raising provocative questions that force teachers to reexamine familiar events and come to see them differently. To understand how insights influence behavior, imagine gaining a new insight into a colleague that makes you more suspicious of him. As a result, you become more guarded in your future interactions with him, more careful about revealing personal information, or perhaps more reluctant to spend time with him. The insight alters your behavior in ways that could not be prescribed by someone else. In fact, if someone else had warned you about this

colleague, you might not have changed your behavior at all because you might not have fully grasped the practical implications of that person's advice. PD programs that rely on insights recognize the importance of teachers' in-the-moment decisions and hope to alter those decisions by changing the way teachers interpret classroom situations in the moment and thus, how they respond to them.

Notice that these three methods of facilitating enactment lie along a continuum in which enactment increasingly depends on teachers' independent judgments. Prescriptions tend to offer universal guidance, allowing teachers very little discretionary judgment. Strategies are also procedurally detailed but their procedures are defined as serving specific purposes and encourage professional judgements as to when they should be used. Insights encourage even more professional judgement, helping teachers learn to "see" situations differently and to make their own decisions about how to respond.

There is yet a fourth approach that moves an additional step further toward teacher autonomy. In this final approach, the PD presents a *body of knowledge* that may not explicitly imply any particular action. By a body of knowledge, I mean knowledge that is organized into a coherent body of interrelated concepts and principles and that can be summarized in books, diagrams, and lectures. PD programs that provide bodies of knowledge often look like traditional university courses with textbooks and syllabi. Bodies of knowledge are inherently passive, and we know very little about how such knowledge rises above its passive "body" status to stimulate any particular teaching action. Thus, when PD offers teachers a body of knowledge, it gives them maximum discretion regarding whether or how teachers would do anything with that knowledge.

I used the frameworks above to define each program's main ideas and methods for facilitating enactment. I developed a set of criteria for connecting program content to the four persistent challenges of teaching and for connecting program pedagogies to each method of facilitating enactment. For example, prescriptions tend to be procedural and universal; strategies also tend to be procedural but they include a purpose and have a multiple-choice quality, in which teachers can select the procedure that seems most appropriate in a given situation. Insights are less explicit and tend to be "discovered" through study groups and discussions rather than didactically. Bodies of knowledge are often defined just as college courses are. When research articles did not provide adequate information about the programs themselves, I went on line and sought further literature about the programs.

I also categorized details of research method, especially with respect to how teachers were allocated to particular groups. When reports were not clear, I contacted authors directly and sought out clarification. With respect to how teachers were assigned to groups, I made a summary judgement as to whether the comparisons ensured comparable motivation to learn and rejected studies that did not.

Results and Interpretation

Distribution of Studies

Table 1 shows the distribution of PD research effort addressing different teaching challenges and different methods of facilitating enactment. Many cells have very few studies within them. Furthermore, studies within a given cell can also

TABLE 1

Distribution of programs across teaching problems and methods of facilitating enactment

	Enactment facilitated by				
Teaching problem	Prescription	Strategy	Insight	Knowledge	Total
Portraying curricular content	5	4	3	3	15
Containing behavior	2	0	0	0	2
Enlisting participation	0	4	1	0	5
Exposing student thinking	1	2	3	0	6
Total	8	10	7	3	28

differ in duration (1 year vs. 2 or 3 years), outcome measures (M1 or M2), and study design (volunteer vs. mandatory assignment). Because each of these variations also has an impact on estimates of program effects, I do not aggregate studies in any way, but instead present their outcomes graphically. This enables readers to consider each potential competing hypothesis when interpreting differences in program effects. I begin with the largest group, which includes programs offering ideas to help teachers portray curriculum content.

Portraying Curriculum Content

Table 2 lists 15 studies whose programs offer ideas about portraying curriculum content, and Figure 1 graphically displays their effects on student achievement. Figure 1 includes a lot of information and requires close inspection. Take a moment to study it. Each icon represents a particular program. They are arrayed along a vertical scale that represents the size of their effects on student achievement. The simplest reading is this: Programs at the top of the graph were more effective than those at the bottom. However, the icons also carry information about the programs themselves, thus allowing for other kinds of comparisons. Larger icons represent larger samples of teacher participants (hence perhaps more reliable estimates), and darker icons indicate more time spent with teachers, with time being a proxy for program intensity. The pattern of light and dark icons invites hypotheses about the relationship between program intensity and program outcomes. In addition, round icons represent M1 outcome measures and square icons represent M2 measures so that readers can compare findings on either type of outcome. Finally, when studies provide findings from both M1 and M2 measures, or provide findings from multiple years, their icons are grouped together or are connected by a line so that readers can see that they come from the same program.

Consider first the vertical arrangement of the icons. Most reside between roughly .00 and .20. This finding may surprise some readers; it has become popular to define effect sizes with terms like *small* or *large*, and effects in the area of .2 or .3 are generally considered small. Using this tradition, almost all PD programs had small effects, regardless of the ideas they offered, how much time they spent with teachers, or how they facilitated enactment of their ideas. But do not

TABLE 2

Programs helping teachers portray curricular content

Citation	Main idea	Facilitated by	Comparison	PD contact hours	Study duration	Number of teachers	M1 outcomes	M2 outcomes
Cole (1992)	Mississippi Competencies	Prescription	No program	24 hours	1 year	12 el	SAT reading SAT mathematics SAT language. arts	
Sloan (1993)	Direct Instruction	Prescription	No program	5 hours	1 year	10 el	CTBS reading CTBS math CTBS science CTBS social studies	
Glazerman et al. (2008); Glazerman et al. (2010); Isenberg et al. (2009)	Comprehensive Induction	Prescription, some strategy	Regular induction	51 hours/ year	3 years	517	Reading Writing	
Borman et al. (2008)	Science Immersion	Prescription	Curriculum materials alone	30 hours	1 year	272 el	District unit science tests	
Garet et al. (2008)	LETRS Coaching	Prescription plus knowledge	Business as usual	48 hours	2 years	270, 250 el	District reading tests	
Penuel et al. (2011)	ESBD (Earth Science by Design)	Strategy	Curriculum materials only	84 hours	1 year	53 sec		Content standards
Matsumura, Garnier, Junker, Resnick, & Bickel (2009)	Questioning the Author	Strategy	Regular literacy coaching	72 hours	2 years	167	State assessment, reading	

(continued)

TABLE 2 (continued)

Citation	Main idea	Facilitated by	Comparison	PD contact hours	Study duration	Number of teachers	M1 outcomes	M2 outcomes
Matsumura, Garnier, Correnti, Junker, and Bickel (2010)	Questioning the Author	Strategy	Regular literacy coaching	36 hours	1 year	73	State assessment, reading	
Campbell and Malkus (2011)	School Math Coaches	Strategy	No program	30 hours	3 years	418	State math assessment	
Supovitz (2012)	Linking Feedback	Insight	Test data alone	3 hours	1 year	64	District unit tests in math	
Gersten, Dimino, Jayanthi, Kim, and Santoro (2010)	Research Study Group	Insight	Other reading first PD	20 hours	1 year	81 el	Oral vocabulary Reading vocabulary Passage comprehension	
Santagata, Kersting, Givven, and Stigler. (2010)	Lesson Analysis	Insight		54 hours	1 year	51 el		Selected items from district test
Garet et al. (2010); Garet et al. (2011)	Rational Numbers	Knowledge	Business as usual	67 hours Year 1 46 hours Year 2	2 years	195 mid, Year 1 92 Year 2	Math knowledge	
Niess (2005)	Oregon Math–Science Partnership	Knowledge	Business as usual	60	1 year	46 el-mid	State math assessment	

Note. el = elementary school; sec = secondary school; mid = middle school; CTBS = Comprehensive Tests of Basic Skills; PD = professional development; LETRS = Language Essentials for Teachers of Reading and Spelling.

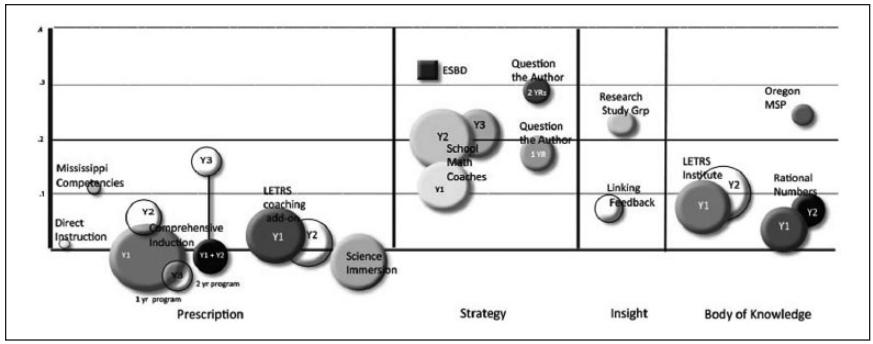

FIGURE 1. *Effectiveness of programs helping teachers portray curricular content.*
Note. Each icon represents a study. The size of the icon is proportional to the number of participating teachers and the darkness is proportional to the amount of time the PD spent with teachers, so that both size and color are indicators of the program's level of effort. When a study followed teachers beyond the treatment year, the later years are represented by hollow icons. Round icons represent M1 measures and square icons represent M2 measures. Small dots layered over an icon mean that the program's effect size was divided by 4 to adjust for class-level units of analysis.

forget that the conventional model of PD is a three-step process: PD alters teachers' knowledge, which in turn alters their practices, which in turn alters student learning. If there is slippage in any one of these steps, we might expect effects to be diminished. Furthermore, when programs use coaches or other intermediaries to work with teachers, they are essentially adding yet another step to this process: They train the coaches, who then work with teachers. The nature of PD, therefore, is such that we cannot rely on conventional definitions of small or large. Indeed, Figure 1 suggests that, for studies of PD, an effect size of .2 could itself be considered rather large, especially when the outcome is a generalized M1 achievement test rather than a more closely aligned M2 test.

Another pattern revealed in Figure 1 has to do with study duration. When programs extend over 2 or more years, multiple icons are displayed, labeled "Y1" and "Y2," to represent effects at the end of each year. When studies follow teachers for a year *beyond the PD year*, their follow-up years are depicted with a white icon, indicating that teachers had no contact with the program during that year, so that we are observing a delayed effect from the treatment. Notice that, in most studies that included such a follow-up year, student achievement was higher at the end of the follow-up year than at the end of the program year. This pattern is consistent with other research suggesting that teachers improve their practices incrementally over time (e.g., Horn, 2010; Huberman, 1994), so that the ultimate effects of PD are likely not completely visible at the end of the program year.

The next pattern revealed in Figure 1 is that the most intensive programs (darker icons) tended to have relatively weaker effects on student learning. Many large, dark circles rest near or even below the baseline of Figure 1. This is a surprising outcome, for these studies represent state-of-the-art research designs. They include larger samples, randomized assignments, and more intensive interventions. But many of these studies also differ from others in that they mandated

participation, a practice that disregards teachers' motivation to learn. Thus, we cannot separate the effect of assignment from the effect of program intensity.

Finally, with respect to the relative merits of different approaches to facilitating enactment, notice that the icons arrayed in Figure 1 form an inverted "U" shape, such that those in the middle of the figure, which relied on strategies or insights to facilitate enactment, had greater effects on student learning relative to those that offered prescriptions or bodies of knowledge. Let us now review the programs individually.

Enactment via Prescription

The leftmost section of Figure 1 presents findings from five programs that relied on prescriptions to facilitate enactment.

- *Mississippi's Competencies* (Cole, 1992) introduced teachers to 14 competencies deemed essential in the early 1990s.
- *Direct Instruction* (Sloan, 1993) introduced teachers to Madeline Hunter's model of instruction, also popular in the early 1990s.
- *Comprehensive Induction* (Glazerman et al., 2008, Glazerman et al., 2010; Isenberg et al., 2009) provided local coaches who observed and guided novice teachers to comply with a rubric-defined set of practices. The researchers examined both a 1-year version and a 2-year version of the program, and in each case followed teachers through a third year.
- LETRS, *or Language Essentials for Teachers of Reading and Spelling* (Garet et al., 2008) shared research-based findings outlined in a National Reading Panel (2000) report. The program provided seminars (called institutes) alone, as well as seminars with coaches who helped teachers implement seminar recommendations. The icon shown in the prescription section of Figure 1 represents the version with coaches.
- *Science Immersion* (Borman et al., 2008) prescribed methods for implementing a new science immersion curriculum in the Los Angeles school system.

These programs differed substantially in the amount of time they spent with teachers. For his *Direct Instruction* program, Sloan (1993) held one meeting with teachers to lay out Hunter's eight elements of a good lesson and to list the specific skills and procedures needed to achieve each element. He then quizzed teachers on what they had learned and provided a second session in which he retaught some of the skills. Even with a second meeting, the total contact time was still just 5 hours. In contrast, the LETRS program provided 48 hours of seminar time and an additional 60 hours of coaching time distributed throughout the year.

To see what highly detailed prescriptive messages look like, consider a sample manual from the *Science Immersion* program. This manual is 206 pages and describes a single fourth-grade unit ("Rot it right," 2006). The manual reminds teachers, in preparation for the unit, to

> ask students to start collecting bottles 4–6 weeks in advance to ensure that you have plenty of bottles to work with. Terraqua Columns require 2-liter plastic bottles. Decomposition Columns require 16 oz. plastic bottles. Make sure that a few extra bottles are available for student groups that encounter a problem while constructing the columns and need to start over. (p. 17)

Each lesson description begins with an enumeration of the steps that need to occur, like this:

> 1. To set the tone for this investigation as an exploration, generate a class discussion and class list about what plants need for growth and development.
>
> 2. Use the Think Aloud technique to model how to refine a wondering into a good scientific investigation. From the students' list about what plants need, form the question—What effect does sunlight have on radish plant growth and development?
>
> 3. Continue the Think Aloud to model assembling the Terraqua Columns using proper experimental procedures, and designing an experiment that has only one factor that is varied.
>
> 4. Have students record and explain their predictions for each set of columns for later reference.
>
> 5. . . . (p. 21)

As this example illustrates, prescriptions are very direct and often very detailed, but Figure 1 suggests that they are not the best way to facilitate enactment of new ideas. However, the three largest studies in this group also relied on mandatory assignments, thus perhaps reducing program effectiveness by removing teachers' motivation to learn.

Enactment via Strategy

The second batch of programs shown in Figure 1 depicts three programs, each of which addressed a different school subject.

- *ESBD, or Earth Science by Design* (Penuel et al., 2011), is based on the broader *Understanding by Design* system (Wiggins & McTighe, 2005). The authors gave teachers a highly scripted step-by step lesson-planning strategy that guided them from long-term goals to specific classroom events, in the hope of helping them become more strategic in their lesson planning.
- *School Math Coaches* (Campbell & Malkus, 2011) provided coaches who collaborated with teachers over a three-year period as they designed their mathematics lessons, thus helping them learn to think more strategically.
- *Questioning the Author* (Matsumura et al., 2010; Matsumura et al., 2013; Matsumura, Garnier, Junker, Resnick, & Bickel 2009) provided coaches to help teachers learn an approach to classroom discussion that might improve students' reading comprehension.

Notice that the *Earth Science by Design* program is represented with a square icon, reminding us that the outcome is an M2 measure rather than an M1 measure. Because M2 measures are more closely aligned with program goals, they often yield larger effects. But all three programs had a greater impact on student

learning than did any of the more prescriptive programs. Notice, too, that two programs in this group have multiple icons because they were both multiyear programs. In both cases, their darker icons depicting later years reflect that the total accumulated time teachers spent across all years of participation.

These two multiyear programs illustrate the problem of focusing on PD design features. Both relied on coaches, as did two programs in the prescriptive group, but these two programs were far more effective than the first two. The concept of coaching is very popular today, and is often recommended as a PD design feature. But coaches in the prescriptive programs used standardized templates to observe and evaluate teachers' practices and to demonstrate recommended practices, whereas coaches in the strategic programs adopted a collaborative, joint problem-solving approach designed to help teachers develop a more strategic approach to their lessons.

Enactment via Insight

Turning to the third section of Figure 1, we see two programs designed to help teachers gain new insights into teaching. There is a third study that also belongs in this group but is not displayed because its report did not provide sufficient information to compute comparable effect sizes.

- The *Research Study Group* (Gersten et al., 2010) introduced teachers to research-based reading strategies, content very similar to that provided by the LETRS program, but it relied on study groups who collectively examined the findings and made their own decisions about how they might design forthcoming lessons in light of these findings.
- *Linking Feedback* (Supovitz, 2012) was designed to add (and link) classroom observation feedback to an already-existing district-wide formative assessment feedback system. In this case, the district had already introduced the assessment system and had formed professional learning communities to examine their data. The only thing the PD program did was add observation feedback in the hope that the linked combination would help teachers generate more or better ideas about what to do next.
- The third program, *Video Lesson Analysis* (Santagata et al., 2010) guided teachers through a process of examining videotaped lessons, conjecturing about how the lesson could have been improved, and ultimately designing and implemented their own lessons for that same content. The program effect could not be calculated in the same metric as the others and so is not displayed in Figure 1. However, its effect was resoundingly negative, making it differ substantially from the others in this group.

All three of these programs relied on some form of professional learning communities, or PLCs, to help teachers gain new insights. Like coaches, PLCs are popular today as a means of engaging teachers in productive discussions about teaching. Yet these three examples of PLC-based PD differed substantially in their effectiveness.

One important difference among these programs had to do with the content they examined. In the most effective program, teachers read research articles about effective practice and discussed the implications of these findings together.

In the middle one, they examined factual information about their own teaching with no suggestions about how to make sense of that feedback. In the third, not shown here, teachers viewed videotapes individually and responded to computerized questions before discussing the videos with colleagues, so their participation may have been relatively more passive. Furthermore, their participation was mandatory, so they might have been less motivated to learn.

Enactment via Bodies of Knowledge

The fourth and rightmost section of Figure 1 displays outcomes from three programs which provided teachers with bodies of knowledge. In these programs, content was presented more didactically, with relatively less attention to implications for enactment.

- *LETRS* (Garet et al., 2008) institutes provided the same content that LETRS coaches did. In this case, teachers attended a series of day-long institutes interspersed throughout the school year, each covering a single research-based topic (e.g., phonemes, phonemic awareness, etc.), and each accompanied by a textbook on that topic. However, teachers in this program did not receive additional guidance from coaches.
- *Rational Numbers* (Garet et al., 2010; Garet et al., 2011) provided intermittent institutes in a format very similar to the LETRS program. Each institute included lectures and overheads interspersed with opportunities for teachers to solve mathematical problems, explain how they solved problems, discuss student misconceptions about these topics, and plan lessons that they would teach later on. The program also included a modest supplemental component to help teachers apply their new knowledge to their classroom instruction.
- The *Oregon MSP, or Mathematics and Science Partnership* (Niess, 2005), used NTCM standards as its organizing framework and offered a 2-week summer institute followed by three semester-length courses in probability and statistics, geometry and measurement, number sense and algebra. There was also an online discussion forum designed to keep teachers' attention focused on this content and to facilitate their use of this new knowledge.

In this section of Figure 1, the two less effective programs also used mandatory assignments, whereas the Oregon MSP study randomly assigned volunteers.

Figure 1 as a Whole

We can now use Figure 1 to compare over a dozen programs that are addressing the same fundamental challenge of teaching, that is, portraying curriculum content in a way that makes it comprehensible to naïve thinkers. But these programs facilitated enactment in different ways. In fact it is now possible to contrast programs that not only addressed the same problem but that also provided the same specific content. Three different programs provided teachers with knowledge about research-based practices for teaching reading and spelling, but they facilitated enactment in different ways. One, the LETRS coaching program, provided prescriptive guidance

inside the classroom. Another, the LETRS institutes, provided a body of knowledge via seminars and textbooks. The third, the Research Study Group, gave each group research reports to think about and discuss as they devised their own lessons. The greater success of this third approach suggests the importance of giving teachers the time and opportunity to make their own sense of new ideas.

But there is another important finding here as well: Notice that the two versions of LETRS also differed in their effectiveness, and that LETRS institutes were *more effective alone* than when they were combined with the coaching component. The prevailing wisdom about PD, based on salient design features, suggests that more is better, but these findings raise questions about when "more" might actually be detrimental. One hypothesis is that prescriptions themselves are inherently less effective, so much so that they have diminished the effect of the institutes by themselves. Another is that the negative effect of mandatory assignments becomes stronger as programs become more intensive. That is, teachers who did not choose to be there in the first place become less and less willing to comply over time.

The pattern of program outcomes depicted in Figure 1 invites at least two kinds of hypotheses for how PD influences practice. One has to do with how programs facilitate enactment of their ideas, the other with how program-assignment methods affect program outcomes. Among all 15 studies displayed in Figure 1, program effectiveness averaged .10, but when studies using mandatory assignments are excluded, the average effect increases to .16.

Addressing Other Persistent Challenges of Teaching

I now turn to PD programs that address other persistent challenges of teaching. These programs are listed in Table 3 and depicted in Figure 2. Figure 2 sorts programs according to which challenge they address, rather than according to how they facilitate enactment of their ideas. In the rightmost section are programs already displayed in Figure 1, but now only studies relying on voluntary assignments are included, so that the studies are more comparable with other studies presented here.

Containing Student Behavior

Two programs offered teachers ideas about containing student behavior.

- *PPA-1, or Process–Product Admonitions* (Anderson et al., 1979), consisted of a single 3-hour meeting with teachers in which the authors describing 22 recommended practices and gave teachers a manual. Then the authors visited half of the participants during the year to document implementation. In essence, they created two treatments groups, one consisting only of PD, and the other consisting of PD combined with observation (but only observation, no feedback). The two versions are connected with a line in Figure 2.
- *PPA-2* (Coladarci & Gage, 1984) provided the same content as the above program but instead of meeting with teachers to review their admonitions, these authors simply mailed them the manual.

TABLE 3

Programs helping teachers contain behavior, enlist participation, or expose student thinking

Citation	Main idea	Enactment facilitated by	Content domain	Comparison	Total PD contact hours	Study duration	Number of teachers	M1 outcomes	M2 outcomes
Containing Behavior									
Anderson, Evertson, and Brophy (1979)	PPA-1, Process–product admonitions	Prescription	Language arts	No program	3 hours	1 year	27 el	Reading readiness Reading achievement	
Coladarci and Gage (1984)	PPA-2, Process–product admonitions	Prescription	Generic	No program	0 hours	1 year	28 el	CTBS reading CTBS mathematics	
Enlisting Participation									
Freiberg, Connell, and Lorentz (2001)	CMCD	Strategy	Math	Math program alone	27 hours	1 year	21 (est)	State assessment mathematics	
Sailors and Price (2010)	Reading Coach	Strategy	Language arts	Summer institute alone	17 hours	1 year	44 el-mid	Standardized reading test	
Greenleaf et al. (2011)	Reading in Science	Strategy	Science	Business as usual	46 hours/year	2 years	105 sec	State assessments in reading, language arts, and biology	
McGill-Franzen, Allington, Yokoi, and Brooks (1999)	Using Books in Kindergarten	Strategy	Language arts	Books alone	38 hours	1 year	18 el	PPVT Ohio Word Recognition	Concepts about print Hearing sounds in words Letter Identification Writing vocabulary
Allen, Pianta, Gregory, Mikami, and Lun (2011)	CLASS	Insight	Generic	Videotaped lessons, no feedback	22 hours	2 years	78 sec	State assessment in teacher's subject	

(continued)

TABLE 3 (continued)

Citation	Main idea	Enactment facilitated by	Content domain	Comparison	Total PD contact hours	Study duration	Number of teachers	M1 outcomes	M2 outcomes
Exposing Student Thinking									
Roth et al. (2011)	Science Story Lines	Strategy and insight	Science	Science content alone	64 hours	1 year	48 el		Photosynthesis Water cycle Electricity Food webs
Roth, Taylor, Wilson, and Landes (2013)	Science Story Lines	Strategy and insight	Science	Science content alone	54 hours	1 year	75 el		Food webs Water cycle
Saxe et al. (2001)	IMA (Integrated Mathematics Assessment)	Strategy	Math	Curriculum alone; study group alone	56 hours	1 year	17 el	Math computation	Math concepts
Carpenter, Fennema, Peterson, Chiang, and Loef (1989)	CGI (Cognitively Guided Instruction)	Insight	Math	No program	80 hours	1 year	39 el	ITBS number facts, computation simple problems, complex problems, advanced problems	Interview
Jacobs, Franke, Carpenter, Levi, and Battey (2007)	Algebraic Reasoning	Insight	Math	Business as usual	35 hours	1 year	103		Relational thinking
Mazzie (2008)	Formative Assessment	Knowledge	Science	No program	38 hours	1 year	21 el	State science assessment	

Note. el = elementary school; sec = secondary school; mid = middle school; PD = professional development; CTBS = Comprehensive Tests of Basic Skills; CMCD = Consistency Management and Cooperative Discipline; PPVT= Peabody Picture Vocabulary Test; ITBS = Iowa Test of Basic skills.

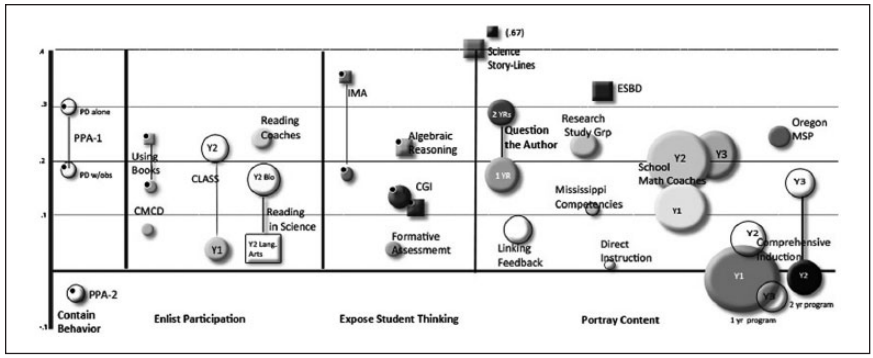

FIGURE 2. *Effectiveness of programs addressing other persistent problems of practice.* *Note.* Each icon represents a study. The size of the icon is proportional to the number of participating teachers, and its darkness is proportional to the amount of time the PD spent with teachers, so that both size and color are indicators of the program's level of effort. When a study followed teachers beyond the treatment year, the later years are represented by hollow icons. Round icons represent M1 measures and square icons represent M2 measures. Small dots layered over an icon mean that the program's effect size was divided by 4 to adjust for class-level units of analysis.

Both programs were prescriptive and both built on early "process–product" research, in which researchers sought relationships between observed teaching practices and student learning. Examples of their prescriptions include "The teacher should use a standard and predictable signal to get the children's attention," or "When call-outs occur, the teacher should remind the child that everyone gets a turn and he must wait his turn to answer."

Notice how different these programs were in their effectiveness, given that their content is exactly the same, and given that all participants were volunteers. These differences introduce the possibility that teachers' participation might also reflect a *social* motivation. When the first authors (Anderson et al., 1979) solicited teachers, they made it clear that they were researchers and that they wanted to test the findings from process–product research. Thus, teachers who agreed to participate did so in order to help the researcher, not because they necessarily believed that they needed to change their own practice or learn anything new. But the social motivation would have been diminished when teachers were observed, and would have disappeared altogether when the admonitions were mailed to them. That this same content could have such different effects across methods of presentation raises important questions for program developers today, when programs are more likely to be mandated, and when multiple programs, rules, and regulations compete for teachers' attention.

Enlisting Participation

Five programs were designed to help teachers enlist student participation. Four of them used strategies to represent their ideas, the fifth relied on insights.

- *CMCD, or Consistency Management and Cooperative Discipline* (Freiberg et al., 2001), emphasized strategies for encouraging sharing, mutual

responsibility, self-discipline, and participation in the formation of classroom rules.

- *Using Books* (McGill-Franzen et al., 1999) offered strategies for kindergarten teachers to engage children with books even though the children could not read, as a way of increasing their motivation to learn. Strategies involved the physical design of the classroom, the use of displays, reading aloud, sorting books into "collections" based on themes, and incorporating literacy activities during play.
- *Reading Coaches* (Sailors & Price, 2010) taught teachers to help children make self-conscious inferences while reading, as a way to foster intellectual engagement. Coaches cotaught some lessons with the teachers, provided demonstration lessons, and provided feedback on the teachers' own practices. Comparison teachers received a common institute but no coaching assistance.
- *Reading in Science* (Greenleaf et al., 2011) introduced secondary biology teachers to the kind of cognitive strategies that are often used by language arts teachers, as a way to help ELL (English-language learners) students who often have difficulty reading science textbooks. The study is unusual in that no data were collected during the program year. So even though the program was very intensive, its icons are white, reflecting the fact that no program activities occurred during the year of data collection.
- *CLASS* (Allen et al., 2011) provided long-distance consultations based on concepts embedded in the Classroom Assessment Scoring System. In the PD, teachers videotaped sample lessons approximately every 2 weeks and sent their tapes to an online "teaching partner." Then the two would talk about the lesson.

The first four of these programs relied on strategies to facilitate enactment. They each provided clear procedural recommendations to teachers but they also provided a clear purpose for their recommendations along with guidance regarding when these strategies would be most appropriate. The fifth program facilitated enactment through insights. Teaching partners were far less directive in their conversations with teachers, and instead used "prompts" to help teachers notice different things in their videotapes. For instance, a "nice work" prompt might say, "You do a nice job letting the students talk. It seems like they are really feeling involved. Why do you think this worked?" A "consider this" prompt might look like this:

> One aspect of Teacher Sensitivity is when you consistently monitor students for cues and when you notice they need extra support or assistance. In this clip, what does the boy in the front row do that shows you that he needs your support at this moment? What criteria did you use to gauge when to move on?

Notice that the teaching partner was not suggesting any specific procedures to teachers, nor explicitly outlining a new strategic goal. Instead, the partner posed questions that might help teachers gain new insights into their own everyday experiences in the classroom.

The three white icons in this group are also worth discussion. All three represent program effects that appeared the year after the program was completed. We can see, for instance, that teachers participating in the CLASS program improved substantially after the program was completed. The Reading in Science icons also represent a follow-up year, but these authors presented no initial-year findings. The Reading in Science program is a model of the kind of PD that experts have been advocating, in that it spent a long time (45 hours) with teachers, sought to ensure that they were actively engaged, and used a lot of classroom artifacts such as videos and examples of student work. Yet it was less effective in its follow-up year than the CLASS program, which spent only about 15 hours with teachers. Once again, these comparisons raise questions about the efficacy of any discrete program design feature.

Exposing Student Thinking

Moving to the third section of Figure 2, we find four studies focused on exposing student thinking.

- *Formative Assessment* (Mazzie, 2008) provided teachers with a 3-credit course on standards-based formative assessment practices. The course covered topics such as performance rubrics, multiple-choice tests, and portfolio assessment but no content about interpreting test scores or revising instruction in response to the scores.
- *IMA*, or *Integrated Mathematics Assessment* (Saxe et al., 2001), was the only program to provide strategies for teachers. Based on a new mathematics curriculum, the program examined the content of each unit, how children thought about that content, and children's motivation to learn that content, and then offered strategies for exposing student thinking during routine classroom activities such as whole-class discussions.
- *Science Story Lines* (Roth et al., 2011; Roth et al., 2013) gave teachers both strategies and insights. The program relied heavily on videotapes of science lessons, encouraging teachers to attend to *story lines* they saw. One story line had to do with how the scientific ideas were connected across lesson segments; the other with how student thinking changed as the lesson progressed. The program also offered clear strategies to help teachers improve their own science lesson story lines and to get better at exposing their students' thinking.
- *CGI, or Cognitively Guided Instruction* (Carpenter et al., 1989), also sought to foster new insights in teachers. Teachers watched videotaped interviews in which children talked about mathematical ideas, so that they could see the students' thought processes. Then they discussed questions such as "How should instruction build on the counting strategies children use to solve simple word problems?" and "How should symbols be linked to children's informal knowledge of addition and subtraction? The authors never suggested any particular procedures or strategies for teaching students, but instead used these conversations to help teachers develop their own ideas.

- *Algebraic Reasoning* (Jacobs et al., 2007) also worked with elementary mathematics teachers, helping them learn about the algebraic relationships that underlay computational arithmetic and to attend to how students understood these relationships. Teachers generated their own lessons but the program encouraged them to pose different kinds of problems to their students and to then bring student responses back to the PD group for examination.

Figure 2 as a Whole

The remainder of Figure 2 displays programs already presented in Figure 1, but it excludes studies that relied on mandated assignments. Figure 2 suggests that guidance about any of these four persistent challenges is equally likely to increase student achievement. This in itself is useful information, in that there is a tendency for critics of education to press for PD that addresses primarily, or only, subject matter knowledge. The pattern displayed in Figure 2 does not suggest that any one problem domain is more likely to help teachers than any other is. On the contrary, helping teachers with any of these persistent problems can lead to gains in student achievement.

Discussion

Most reviews of PD seek to define a list of critical program design features such as program duration, topic, number of contact hours, or types of learning activities (e.g., Blank & de las Alas, 2009; Kennedy, 1998; Sher & O'Reilly, 2009; Timperley et al., 2007). The prevalence of these reviews has led the field as a whole to embrace a set of specific program design features that are presumed to define high-quality PD. Desimone (2009) provided an excellent review of this literature and used it to generate a framework for defining important aspects of PD. Her work was cited by many authors reviewed here, and many authors reviewed here assured readers that their PD programs met these widely advocated design features. Yet this review suggests that program design features may be unreliable predictors of program success.

Probably the most widely referenced requirement for PD is that it should focus on content knowledge (Blank et al., 2008; Desimone, 2009; Yoon et al., 2007). Many authors of PD studies reviewed here (e.g., Greenleaf et al., 2011; Heller et al., 2012; Niess, 2005) cite the importance of content knowledge as a rationale for their programs. Yet the findings presented here suggest that programs addressing any of the four persistent problems of teaching can improve teachers' effectiveness, and, in fact, programs that focused *exclusively* on content knowledge tended to have less effect on student learning. When programs offering content knowledge were successful, the content was subsumed under a broader goal, such as helping teachers learn to expose student thinking.

Another widely recognized design feature is *collective participation* (e.g., Blank et al., 2008; Desimone, Garet, Birman, Porter, & Yoon, 2003; Loucks-Horsley & Matsumoto, 1999; Yoon et al., 2007), and some authors whose studies are reviewed here emphasized the importance of professional learning communities, or PLCs, as a rationale for their PD designs. But learning communities also varied in their effectiveness and one of them, using video-based lesson analysis,

had a negative impact on student learning. Furthermore, at least one study in this review (Saxe et al., 2001) used professional learning communities as the *comparison group* against which to test its own approach to PD. One reason for these surprising findings is that different versions of PLCs differed in important, but less often noticed, ways. In the most effective PLCs, the Research Study Groups (Gersten et al., 2010), teachers were given research findings to think about, and each group had a discussion leader to pose questions and keep conversations on target. In contrast, teachers in the linking study were given factual information about their students' achievement and about their own classroom practices, but were left to their own devices to make sense of that information. The third program using PLCs relied on analysis of videotaped lessons, but much of the teachers' work was done individually as they responded to a series of programmed questions. As researchers, we need to move past the concept of learning communities per se and begin examining the content such groups discuss and the nature of intellectual work they are engaged in.

Another widely mentioned program design feature is program *intensity*, which sometimes refers to the total amount of contact hours with teachers, sometimes to the total span of time over which these hours are distributed, and sometimes to the volume of information transmitted. At least one set of authors reviewed here (Greenleaf et al., 2011) mentioned a long span of time as a rationale for their program design; many others mentioned the relevance of a large number of contact hours. But a glance at the Figures 1 and 2 suggests that more intense programs—those with the darkest icons—were not necessarily more likely to rise to the top of the graphs. Dark icons seem to appear at the bottom of these figures as often, or even more often, than at the top. But other variables, not discussed in the literature, are relevant here. Program intensity appears to be less effective when combined with prescriptive messages, for instance, but more effective when messages provide strategies or insights.

Finally, another widely recognized program design feature is the use of coaches, and again this review shows that coaches vary in their value. But again, the value of coaches seems to depend on how they try to facilitate enactment. Coaches in the LETRS and Comprehensive Induction programs tended to observe and evaluate teachers for how well they complied with an observation rubric, whereas coaches in more effective programs collaborated with teachers on lesson planning, providing a model of strategic planning.

The limitations in popular lists of design features is nowhere more evident than in the results of the earliest study reviewed (Anderson et al., 1979), in which the authors spent only 3 hours with teachers describing a list of process–product research findings. The program is among the most effective displayed in Figure 2. Yet it made no mention of subject matter knowledge, spent very little time with teachers, did not actively engage teachers in any learning activities, did not embed its ideas in the curriculum, school settings or classroom settings, and did not invite collective participation. Teachers participated individually rather than in groups, and were simply given a list of general guidelines for their practice. Yet the program yielded strong effects on student learning. One reason for this surprising outcome may have to do with the study context. In 1979, PD was not nearly as ubiquitous as it now is. The program likely offered a unique experience for participating teachers and it did not compete with other initiatives for their attention.

But another important distinction is that the authors treated the teachers more as colleagues whose role was to help the researchers test this new model of instruction, rather than as teachers whose practices needed improvement. Thus, participation was in part socially motivated.

Education research is at a stage in which we have strong theories of *student* learning, but we do not have well-developed ideas about *teacher* learning, nor about how to help teachers incorporate new ideas into their ongoing systems of practice. This disjuncture yields programs such as the Los Angeles Science Immersion program which aims to actively immerse students in scientific activities but at the same time inundates teachers with volumes of prescriptive details about how they should immerse their students in science. Why would we expect these detailed prescriptions to work for teachers if we do not believe that they work for students?

We also need to pay more attention to the people who provide PD. We have an extensive literature on the kind of knowledge teachers need for teaching, including constructs such as with-it-ness (Kounin, 1970), pedagogical content knowledge (Shulman, 1987), and mathematical knowledge for teaching (Hill, Rowan, & Ball, 2005) but we do not examine the knowledge needed by professional developers, nor do we have a language with which to characterize the environments of their PD "classrooms." Many of the more effective programs reviewed here were offered by individuals or groups who had long histories of working with teachers, were very familiar with teachers and with the problems they face, and based their programs on their own personal experience and expertise. Many of the less effective programs were large-scale programs that relied on intermediaries—coaches or small group facilitators who were hired specifically for the study, and whose familiarity with teaching, or more importantly, with teacher learning, may have been limited. There is little discussion in the literature about the nature of PD expertise, how PD providers are selected, how they are prepared for their work, or how their efficacy is assessed. These topics need to become part of our discussion as we generate and test our PD theories of action.

With respect to research design, this review introduces new questions about the role of motivation in PD. Mandated PD creates a problem for PD developers, which is analogous to the problem teachers face: *Attendance is mandatory but learning is not*. Among the entire array of studies reviewed here, the average effect from studies that assigned volunteers was .16 on M1 measures, whereas the average effect among studies using mandated assignments was .03. Studies that scrupulously comply with the WWC assign teachers who have not indicated any interest in PD. These studies cannot benefit from teachers' motivation to learn and are not good tests of the potential of the PD programs.

In addition to acknowledging the role of motivation in learning, we also need research designs that acknowledge the slow and incremental way in which teachers incorporate new ideas into their ongoing practices (see, e.g., Huberman, 1994). Studies that are coterminous with the PD itself cannot tell us whether teachers merely comply with program recommendations as long as they have to, whether they continue building on the program's ideas over time, or whether, perhaps, they revise the advice so severely that its original meaning is lost. The differences we see here between program effects at the end of the PD versus program effects a

year later provide a strong argument for researchers to follow teachers beyond the end of the PD, and I would urge researchers to monitoring student learning for 1 or 2 years beyond the close of the PD itself.

Finally, we need to ask hard questions about programs that have *negative* effects on teachers. It is certainly possible for a program to fail, but failure should yield a *null* effect, not a negative effect. After looking closely at the programs reviewed here, I suspect that negative effects arise from negative emotional responses—perhaps resistance or resentment toward the program's demands.

If we can tie our research designs and our PD models more closely to underlying theories of teacher motivation and teacher learning, we will learn more from our studies. We need to replace our current conception of "good" PD as comprising a collection of particular design features with a conception that is based on more nuanced understanding of what teachers do, what motivates them, and how they learn and grow. We also need to reconceptualize teachers as people with their own motivations and interests. The differences shown here among PD methods of facilitating enactment strongly suggest the importance of intellectually engaging teachers with PD content, rather than simply presenting prescriptions or presenting bodies of knowledge. Furthermore, the differences in program effectiveness when studies compared groups of volunteers as opposed to groups of nonvolunteers remind us of the role of teachers' own volition in improving their practices. Future research should attend more to how PD programs motivate teachers, how they intellectually engage teachers, and to whether programs are meaningful to teachers themselves. This is especially important in an era in which teachers receive numerous messages about what they should be doing and in which these messages compete for teachers' attention. We need to ensure that PD promotes real learning rather than merely adding more noise to their working environment.

Note

The author wishes to acknowledge helpful feedback from many colleagues. They are obviously not responsible for the article as it currently stands but were all generous, thoughtful, and constructive in their responses to earlier drafts. These colleagues include Laura Desimone, University of Pennsylvania; Michael Garet, American Institutes for Research; Catherine Lewis, Mills College; Kathy Roth, BSCS, Colorado Springs; Tanya Wright, Michigan State University; and Meng-Jia Wu, Loyola University of Chicago.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*, 1034–1037. doi:10.1126/science.1207998

*Anderson, L. M., Evertson, C. M., & Brophy, J. E. (1979). An experimental study of effective teaching in first-grade reading groups. *Elementary School Journal*, *79*, 193–223. doi:10.1086/461151

Blank, R. K., & de las Alas, N. (2009, June). *Effects of teacher professional development on gains in student achievement: How meta-analysis provides scientific evidence useful to educational leaders*. Washington, DC: Council of Chief State School

Officers. Retrieved from http://www.ccsso.org/Documents/2009/Effects_of_Teacher_Professional_2009.pdf

Blank, R. K., de las Alas, N., & Smith, C. (2008, January). *Does teacher professional development have effects on teaching and learning? Evaluation findings from programs in 14 states*. Washington, DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Resources/Publications/Does_Teacher_Professional_Development_Have_Effects_on_Teaching_and_Learning_Analysis_of_Evaluation_Findings_from_Programs_for_Mathematics_and_Science_Teachers_in_14_States.html

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

*Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness*, *1*, 237–264. doi:10.1080/19345740802328273

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington DC: National Academy Press.

*Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *Elementary School Journal*, *111*, 430–454. doi:10.1086/657654

Campbell, P. F., & Rowan, T. E. (1995). *Project IMPACT: Increasing the mathematical power of all children and teachers*. College Park: University of Maryland.

*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. doi:10.3102/00028312026004499

Charters, W. W., & Waples, D. (1929). *Commonwealth teacher training study*. Chicago, IL: University of Chicago Press.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 255–296). New York, NY: Macmillan.

Cohen, D. K. (2011). *Teaching and its predicaments*. Cambridge, MA: Harvard University Press.

*Coladarci, T., & Gage, N. L. (1984). Effects of a minimal intervention on teacher behavior and student achievement. *American Educational Research Journal*, *21*, 539-555. doi:10.3102/00028312021003539

*Cole, D. C. (1992). *The effects of a one-year staff development program on the achievement of fourth-grade students* (Doctoral dissertation). Retrieved from University of Mississippi Database. (UMI No. 9232258)

Cusick, P. A. (1983). *The egalitarian ideal and the American high school: Studies of three schools*. New York, NY: Longman.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, 181–199. doi:10.3102/0013189X08331140

Desimone, L. M., Garet, M. S., Birman, B. F., Porter, A., & Yoon, K. S. (2003). Improving teachers' in-service professional development in mathematics and science: The role of postsecondary institutions. *Education Policy*, *17*, 613–649. doi:10.1177/0895904803256791

Doyle, W. (1986). Content representation in teachers' definitions of academic work. *Journal of Curriculum Studies*, *18*, 365–379. doi:10.1080/0022027860180402

*Duffy, G. G., Roehler, L. R., Meloth, M. S., Vavrus, L. G., Book, C., Putnam, J., & Wesselman, R. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects. *Reading Research Quarterly*, *21*, 237–252. doi:10.2307/747707

*Freiberg, H. J., Connell, M. L., & Lorentz, J. (2001). Effects of consistency management on student mathematics achievement in seven Chapter I elementary schools. *Journal of Education for Students Placed at Risk*, *6*, 249–270. doi:10.1207/S15327671ESPR0603_6

Fuchs, L. S. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal*, *28*, 617–641. doi:10.3102/00028312028003617

*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*. Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/pdf/20084030.pdf

*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*. Washington, DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20114024

*Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., . . .Hurlburt, S. (2010). *Middle school mathematics professional development impact study: Findings after the first year of implementation*. Washington, DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20104009/

*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. doi:10.3102/0002831209361208

*Glazerman, S., Dolfin, S., Bleeker, M., Johnson, A., Isenberg, E., Lugo-Gil, J., . . . Britton, E. (2008, October). *Impacts of comprehensive teacher induction: Results from the first year of a randomized controlled study* (NCEE 2009-4035). Washington, DC: Institute for Education Sciences. Retrieved from http://ies.ed.gov/ncee/pdf/20094035.pdf

*Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010, June). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study*. Washington, DC: National Center for Education Evaluation. Retrieved from https://ies.ed.gov/ncee/pubs/20104027/

*Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., . . . Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of Reading Apprenticeship professional development. *American Educational Research Journal*, *48*, 647–717. doi:10.3102/0002831210384839

Heller, J. I., Dahler, 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*, 333–362. doi:10.1002/tea.21004

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K., Hollingsworth, H., Jacobs, J., . . .Stigler, J. (2003). *Teaching mathematics in seven countries: TIMSS 1999 video study*. Washington, DC: National Center for Educational Statistics. Retrieved from http://timssvideo.com/timss-video-study

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*, 371–406. doi:10.3102/00028312042002371

Hirsh, S. A., & Hord, S. M. (2008). Role of professional learning in advanced quality teaching and student learning. In T. L. Good (Ed.), *21st century eEducation: A reference handbook* (Vol. 2, pp. 337–347). Los Angeles, CA: Sage.

Horn, I. S. (2010). Teaching replays, teaching rehearsals, and re-visions of practice: Learning from colleagues in a mathematics teacher community. *Teachers College Record*, *112*, 225–259.

Huberman, M. (1994). *The lives of teachers*. New York, NY: Teachers College Press.

*Isenberg, E., Glazerman, S., Bleeker, M., Johnson, A., Lugo-Gil, J., Grider, M., & Dolfin, S. (2009, August). *Impacts of comprehensive teacher induction: Results from the second year of a randomized controlled study*. Washington, DC: Institute for Education Sciences. Retrieved from http://ies.ed.gov/ncee/pubs/20094072/pdf/20094072.pdf

*Jacobs, V. R., Franke, M. L., Carpenter, T., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education*, *38*, 258–288.

Kennedy, M. M. (1998). *Form and substance in inservice teacher education*. Madison: University of Wisconsin National Institute for Science Education. Retrieved from https://www.msu.edu/~mkennedy/publications/valuePD.html

Kennedy, M. M. (1999). The role of preservice teacher education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 54–85). San Francisco, CA: Jossey Bass.

Kennedy, M. M. (2005). *Inside teaching: How classroom life undermines reform*. Cambridge, MA: Harvard University Press.

Kennedy, M. M. (2016). Parsing the practice of teaching. *Journal of Teacher Education 67*, 6-17. doi:10.1177/0022487115614617

Kounin, J. (1970). *Discipline and group management in classrooms*. New York, NY: Holt, Rinehart & Winston.

Lampert, M. (1988). What can research on teacher education tell us about improving the quality of mathematics education? *Teaching and Teacher Education*, *4*, 157–170. doi:10.1016/0742-051X(88)90015-7

Loucks-Horsley, S., & Matsumoto, C. (1999). Research on professional development for teachers of mathematics and science: The state of the scene. *School Science and Mathematics*, *99*, 258–271. doi:10.1111/j.1949-8594.1999.tb17484.x

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*, 41–53. doi:10.1002/tea.3660280105

*Matsumura, L. C., Garnier, H. E., Correnti, R., Junker, B., & Bickel, D. D. (2010). Investigating the effectiveness of a comprehensive literacy coaching program in schools with high teacher mobility. *Elementary School Journal*, *111*, 35–62. doi:10.1086/653469

*Matsumura, L. C., Garnier, H. E., Junker, B., Resnick, L., & Bickel, D. B. (2009, March). *The influence of Content-Focused Coaching® on reading comprehension instruction and student achievement*. Paper presented at the Society for Research on Educational Effectiveness, Washington, DC.

Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction*, *25*, 35–48. doi:10.1016/j.learninstruc.2012.11.001

*Mazzie, D. D. (2008). *The effects of professional development related to classroom assessment on student achievement in science* (Doctoral dissertation). Retrieved from University of South Carolina database. (UMI No. 3321420)

McCutchen, D., Abbott, R. D., Green, L. B., & Beretvas, S. N. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities*, *35*, 69–86. doi:10.1177/002221940203500106

*McGill-Franzen, A., Allington, R. L., Yokoi, L., & Brooks, G. (1999). Putting books in the classroom seems necessary but not sufficient. *Journal of Educational Research*, *93*, 67–74. doi:10.1080/00220679909597631

McNeil, L. M. (1985). *Contradictions of control: School structure and school knowledge*. London, England: Routledge & Kegan Paul.

Metz, M. (1993). Teachers' ultimate dependence on their students. In J. Little & M. McLaughlin (Eds.), *Teachers' work: Individuals, colleagues and context* (pp. 104–136). New York, NY: Teachers College Press.

Mitchell, L., & Cubey, P. (2003). *Characteristics of professional development linked to enhanced pedagogy and children's learning in early childhood settings*. Wellington, New Zealand: Ministry of Education.

National Reading Panel. (2000). *Teaching children to read: An evidence based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Washington, DC: National Institute of Child Health and Human Development. Retrieved from https://www.nichd.nih.gov/publications/pubs/nrp/documents/report.pdf

Newman, D., Finney, P. B., Bell, S., Turner, H., Jaciw, A., Zacamy, J. L., & Gould, L. F. (2012). *Evaluation of the effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI)*. Washington, DC: Institute for Education Sciences.

*Niess, M. (2005). *Oregon ESEA Title IIB MSP*. Corvallis: Central Oregon Consortium.

Opfer, D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, *81*, 376–407. doi:10.3102/0034654311413609

*Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth science: A comparison of three professional development programs. *American Educational Research Journal*, *48*, 996–1025. doi:10.3102/0002831211410864

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., . . .Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*, 833–878. doi:10.3102/0002831210367426

**Rot it right: The cycling of matter and the transfer of energy. 4th Grade Science Immersion Unit*. (2006, September). System Wide Change for All Learners and Educators (SCALE) Report. Madison: University of Wisconsin–Madison. Retrieved from www.scalemsp.org

Roth, K. J., Druker, S. L., Garnier, H. E., Lemmens, M., Chen, C., Kawanaka, T., . . . Gallimore, R. (2006). *Teaching science in five countries: Results from the TIMSS 1999 video study*. Washington, DC: U.S. Department of Education National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2006/2006011.pdf

*Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Video-based lesson analysis: Effective science PD for teacher and student learning. *Journal for Research in Science Teaching*, *48*, 117–148. doi:10.1002/tea.20408

*Roth, K. J., Taylor, J., Wilson, C., & Landes, N. M. (2013, April 9). *Scale-up study of a videocase-based lesson analysis PD program: Teacher and student science content learning*. Paper presented at the National Association of Research on Science Teaching, Rio Grande, Puerto Rico.

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

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. P. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*, 369–393. doi:10.1002/tea.10027

*Sailors, M., & Price, L. R. (2010). Professional development that supports the teaching of cognitive reading strategy instruction. *Elementary School Journal*, *110*, 301–322. doi:10.1086/648980

*Santagata, R., Kersting, N., Givven, K. B., & Stigler, J. W. (2010). 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. doi:10.1080/19345747.2010.498562

*Saxe, G. B., Gearhart, M., & Nasir, N. I. S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. *Journal of Mathematics Teacher Education*, *4*, 55–79. doi:10.1023/A:1009935100676

Sedlak, M., Wheeler, C., Pullin, D. C., & Cusick, P. A. (1986). *Selling students short: Classroom bargains and academic reform in the American high school*. New York, NY: Teachers College Press.

Sher, L., & O'Reilly, F. E. (2009). Professional development for K-12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness*, *2*, 209–249. doi:10.1080/19345740802641527

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

Siegle, D., & McCoach, D. B. (2007). Self-efficacy through teacher training. *Journal of Advanced Academics*, *18*, 278–312.

Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*, *51*, 870–901. doi:10.1080/10824669.2013.862095

*Sloan, H. A. (1993). *Direct instruction in fourth and fifth-grade classrooms* (Doctoral dissertation). Retrieved from Purdue University database. (UMI No. 9334424)

Stigler, J. W., & Perry, M. (1988). Mathematics learning in Japanese, Chinese, and American classrooms. In G. B. Saxe & M. Gearhart (Eds.), *Children's mathematics: New directions for child development* (No. 41, pp. 27–54). San Francisco, CA: Jossey Bass.

*Supovitz, J. (2012). *The Linking Study—First year results: A report of the first year effects of an experimental study of the impact of feedback to teachers on teaching and learning*. Philadelphia, PA: University of Pennsylvania Center for Policy Research in Education. Retrieved from http://www.cpre.org/linking-study-first-year-results

Sztajn, P., Campbell, M. P., & Yoon, K. S. (2011). Conceptualizing professional development in mathematics: Elements of a model. *PNA*, *5*(3), 83–92.

Tienken, C. H., & Achilles, C. M. (2003). Changing teacher behavior and improving student writing achievement. *Planning and Change*, *34*, 153–168.

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration (BES)*. Auckland, New Zealand: University of Auckland. Retrieved from http://educationcounts.edcentre.govt.nz/goto/BES

Wei, R. C., Darling-Hammond, L., & Adamson, F. (2010). *Professional development in the United States: Trends and challenges*. Dallas, TX: National Staff Development Council. Retrieved from http://learningforward.org/docs/pdf/nsdcstudytechnicalreport2010.pdf?sfvrsn=0

What Works Clearinghouse. (2013). *Procedures and standards handbook (Version 3.0)*. Washington, DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

Wideen, M., Mayer-Smith, J., & Moon, B. (1998). A critical analysis of the research on learning to teach: Making the case for an ecological perspective on inquiry. *Review of Educational Research*, *68*, 130–178. doi:10.3102/00346543068002130

Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Spapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Washington, DC: Institute for Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

Author

MARY M. KENNEDY (Department of Teacher Education, Michigan State University, 708 Applegate Lane, East Lansing, MI 48823; e-mail: *mkennedy@msu.edu*) has a long history of scholarship focused on defining teaching quality and identifying the factors that most influence teaching quality. She is a fellow in the American Educational Research Association. Her book *Inside Teaching: How Classroom Life Undermines Reform* (2005) addresses the influence of school context on the quality of teaching practices and shows how local circumstances make it difficult for teachers to live up to reform expectations.