Chapter XX

How We Learn About Teacher Learning

Mary M. Kennedy
Michigan State University

This chapter examines research on professional development, or PD, focusing specifically on underlying assumptions about the nature of teaching and the nature of teacher learning. It examines PD programs according to their assumptions about what teachers need to learn, and it examines PD studies according to how and when they expect to see evidence of teacher learning. The chapter seeks to provide a broad view of how we think about teaching and teacher learning and to examine our underlying assumptions both about teaching and about how PD is expected to improve teaching. With respect to program effectiveness, the chapter raises questions about the extent to which effective PD programs can be replicated; with respect to our study designs, it raises questions about how teacher learning occurs and when and how we should expect to see program effects on teachers' practices. The chapter also offers some suggestions for future research design.

Human beings have taught one another for centuries, and for most of that time everyone invented their own approaches to teaching, without the guidance of mentors, administrators, teacher educators, or professional developers. Today, teachers receive guidance from almost every corner. They are formally certified to teach, and once certified, they continue to take additional courses, called *professional development*, or PD, throughout their teaching careers. In addition, states and school districts also regulate many aspects of their work through performance appraisals and student assessments.

This chapter addresses a specific portion of guidance called professional development, or PD. Literature on PD has grown substantially over time, and standards for research have also changed. Twenty years ago, I reviewed studies of PD effectiveness within math and science education (Kennedy, 1998), limiting my review to studies that provided evidence of student achievement and that included a comparison group. I found only 12 such studies, most of which are not acceptable by today's

Review of Research in Education
Month 201X, Vol. XX, pp. 1–25
DOI: 10.3102/0091732X19838970
Article reuse guidelines: sagepub.com/journals-permissions

standards. Some had very small samples, some did not randomly assign teachers to groups, and some provided instructional materials as well as PD, so that the effects of the PD were confounded with the effects of the materials. Since then, the literature has grown substantially, so that we are able to raise our standards for what counts as a "good study." Two years ago, I reviewed 28 studies of PD (Kennedy, 2016), all of which used random assignment. Since then, even more such studies have been published. For this chapter, I now raise my standards again and remove studies that were based on fewer than 20 teachers.[1]

PD studies represent our way, as researchers, of learning about teacher learning. We generate hypotheses about what good teaching consists of, about what teachers need to learn in order to do good teaching, or about what kind of activities or experiences provoke learning in teachers. Then we try different kinds of interventions to see how they work. It is not a perfect system, for every PD study simultaneously involves all three of these types of hypotheses: what teachers need to learn, how they learn, and how we will know whether they have learned enough. Thus, a given study of PD can fail if any one of these hypotheses is wrong, and we may not know where our error is. Furthermore, teachers themselves may learn about teaching independently, in ways we don't see. They take formal courses, they read things, they ruminate about their own experiences, and they seek advice from colleagues. They may even get a brainstorm about their teaching while watching a movie.

My aim in this chapter is to examine our existing oeuvre of experimental research on PD both from the standpoint of what we have learned about teacher learning and from the standpoint of what we have learned about *how to learn* about teacher learning—that is, how to design informative studies. I begin with a brief overview of how we think about teaching as a phenomenon.

FIRST IMPRESSIONS OF TEACHING

Teaching is, among other things, a cultural activity. We have all spent thousands of hours observing teachers and participating in classroom activities, and we have all formed a variety of different and sometimes contradictory thoughts about teaching. Here I offer five observations about our current understandings of teaching.

Our Learning Begins in Childhood

Learning about teaching is different from learning about any other occupation, in that our learning begins when we are children. As we watch our own teachers, we develop ideas about what they are doing, why they are doing it, what effect their work has on us, and so forth. The sociologist Dan Lortie (1975) referred to this extended period as an "apprenticeship of observation" and pointed out that this kind of occupational familiarity is unique to the profession of teaching. All of us—those who become teachers, those who become education researchers, and everyone else—have spent roughly 12,000 hours watching teachers through our child-eyes, developing our own conceptions about what the job entails and what makes some teachers better than others.

But the ideas we form about teaching are *naive* in the sense that they are formed without any awareness of what really causes events to turn out as they do. Just as a child might form the naive conception that the sun circles around the earth, she/he might form the conception that teaching practice comes naturally, or is effortless, because teachers always appear to know what to do. Moreover, we remain confident in our judgments that one teacher is "better" than another.

This is an important preface to any discussion about learning about teaching because our impressions could be wrong. We may be aware of the effect of a teacher's actions but not what its purpose was. We see their actions but not their thoughts, their goals, their motives, their frustrations. Moreover, we don't see *what they see*, from their vantage point at the front of the classroom and from their vantage point of trying to lead the class in a particular direction. This lack of awareness, in turn, can lead us to think that teaching practices come naturally or that the decisions about "what to do next" are always self-evident in the moment.

I became especially aware of the difference between an observer's view and a teacher's view in a study of teachers' in-the-moment decision making (Kennedy, 2005, 2010b). When I asked teachers about discrete actions they took during a lesson, they nearly always referred to something they saw at that moment. Teachers would say "I could see that Billy was about to jump out of his seat," or "I realized I didn't have enough handouts to go around," or "Juan rarely speaks and I wanted to encourage him." These conversations reveal a highly contingent aspect of teaching that is quite different from our naive conceptions of teachers as entirely self-directed and always knowing what to do next.

As Researchers, We Like Idealized Models of Teaching

In the 1970s, a federal program called *Follow Through* supported the development and field-testing of different models of teaching (McDaniels, 1975) that exemplified different teaching ideals. One model, for instance, was based on the research of the French psychologist, Jean Piaget, while another was based on behaviorist theories of learning, a third on the concept of open classrooms, and a fourth on the concept of a *Responsive Environment*. Each model developer was called a sponsor, and sponsors were funded to train teachers in specific schools to implement their models. Eventually over a dozen such models were developed and field-tested in schools throughout the country.

Models are useful to researchers because they provide us with a nomenclature that can be used in our research to distinguish among teachers. Researchers today continue to design, study, and evaluate different models of teaching. Some models derive from naive conceptions of teaching, some from theories of student learning, and some from empirical evidence of relationships between specific teaching practices and student learning. But many still embrace the naive view of the teacher as always in full control of the classroom, still failing to recognize the contingent nature of teaching. This view of teaching practice as entirely in the teachers' control leads to what social psychologists call *Attribution Error*, a tendency to assume that the behaviors we observe in

others are caused entirely by their own character, not by the situations they are confronting.

We Are Guilty of Attribution Error

Since we tend to assume teaching comes naturally, and that teachers are entirely in control of events in their classrooms, we also assume that whatever behaviors we see are purposeful, rather than spurious responses to events. Here is an example of an attribution error I made several years ago when observing a fifth-grade mathematics teacher, Ms. Katlaski. It was the first period of the day, and she had been on hall duty that morning. When the bell rang, she entered the room and spent less than a minute looking at the text to remind herself what the lesson was about. It was about multiplying whole numbers with fractions. Students had previously learned how to multiply whole numbers with each other, and how to multiply fractions with each other. Today they would learn to solve a problem involving both a whole number and a fraction: $9 \times 2/3$. Her plan had been to show them that they could convert the 9 into 9/1, so that they could then use computation strategies they had already learned. To open the lesson, Katlaski asked the rhetorical question of how to convert 9 into a fraction. She was not really expecting an answer, but in this case, someone called out, Multiply 9 by 4/4. The student's proposal was technically correct, but it would be mathematically more difficult to solve if you are 9 years old. Katlaski's solution, converting the 9 to 9/1, would have yielded this computation:

$$9 / 1 \times 2 / 3 = 18 / 3 = 6.$$

The student's solution would have yielded this computation:

$$36 / 4 \times 2 / 3 = 72 / 12 = 6.$$

Katlaski knew that the student's proposal would be too complicated for her students to follow, so she immediately faced a dilemma: accept the student's solution and solve the problem on the board, even if most students couldn't follow it, or reject the student's solution. In a fluster, she said, "No, that won't work."

Katlaski's behavior would imply that she did not know her mathematics, but her real problem was a logistical one of how to respond to a proposal that would be difficult for her students to follow (Kennedy, 2010a). The term "attribution error" refers to this tendency to attribute the actions of others to stable personal traits rather than to the situations in which they find themselves. In Katlaski's case, the situation presented something she wasn't ready for, and an unknowing observer could easily attribute that error to a lack of sufficient content knowledge. Because we are all vulnerable to the assumption that teachers always know what to do, we are especially guilty, even as grown-ups and even as researchers, of attributing teachers' actions to their content knowledge or their character traits rather than to the situations they face.

We Expect PD to Solve All the Problems We See

In the past several decades, the number and variety of PD programs developed for teachers, and often required of them, have continuously increased. By 2001, teacher unions were adding PD requirements into their contracts, typically stipulating the number of hours of PD they should be taking each year (see, e.g., Bredeson, 2001). One recent study (TNTP, 2015) found some school districts that spent an average of $18,000 *per year per teacher* on PD. Yet education literature is replete with articles about the need for even more of it. A Web search for "professional development for teachers" yields dozens of sites advocating more or better PD for teachers. And as advocacy for PD has increased, so has research on PD.

My aim in this paper is to examine the research on PD from two perspectives: First, what have we learned about the benefits of these PD programs for teachers and students? Second, what have we learned about *how to study* the benefits of PD for teachers? For this analysis, I rely on a population of 29 PD studies that (a) include a minimum of 20 teachers, (b) include at least one measure of student achievement as an outcome, (c) include a comparison group, and (d) rely on some form of random assignment to allocate teachers to treatment and comparison conditions.[2] This population of studies is large enough to enable us to examine patterns in their findings and perhaps generate some useful hypotheses about when, why, or how PD can be helpful. All of these studies involve formal PD programs. That is, people from somewhere outside the school held scheduled meetings with teachers in order to alter teachers' practices in some specific way.

All the studies in this review have a comparison group, but I also rely here on a comparison *study*, in which the treatment of interest involves pairing relatively weaker teachers with other teachers in their own schools.

An Alternative to PD: Let Teachers Help Each Other

When researchers bring their idealized models of teaching into schools, there are at least two ways they can go wrong. They could be wrong about the effectiveness of their idealized model of teaching, or their PD could fail to address the myriad contingencies of teaching and thereby provide no benefit to teachers. A useful contrast to formal PD, therefore, is the system teachers often rely on informally, which is to seek guidance from one another and share tips about how to handle various contingencies. I found one study (Papay, Taylor, Tyler, & Laski, 2016) that pursues this idea. These researchers asked school principals to pair teachers whose practices fell below district standards with other teachers whose performances were higher. Although there was an official "curriculum" for this peer-to-peer support system, which was based on the district's performance assessment system, this program itself was very flexible: Principals were at liberty to decide which teachers from one group would be paired with which teacher from the other group. Then, mentors were free to do or say whatever they wanted to their protégés, and protégés were free to accept or reject their mentors' advice. The researchers hoped that the mentor teachers would discuss

specific practices in which protégés were known to be less effective, but there were no rules regarding the content or format of these conversations. One teacher might present a specific behavior as a requirement, to be done at least twice every day, while another might cast it as useful in specific types of situations.

This program provides a useful alternative to conventional PD with conventional curricula: The program had no cost, no formal schedule, and no uniform curriculum. Formal PD programs have all these things—standards, goals, models of good practice, admonitions, and so forth—but they also have substantial cost and take up a lot of teachers' time. Since virtually all approaches to PD will be more expensive, we should expect them to demonstrate greater value than this simple "bootstrap" approach.

HOW DO WE MEASURE THE BENEFITS OF PD?

Throughout the history of education research, we have estimated the benefits of experimental programs with tests of statistical significance. These tests tell us the likelihood of achieving an outcome by chance, but they do not help us estimate the practical relevance of the effect itself. Effect sizes, on the other hand, allow us to place all differences between groups on the same standardized scale, typically ranging from –1 to +1, regardless of what outcome measure is used and what research design is used. An early advocate for the use of effect sizes, Jacob Cohen (1988) offered some rough guidelines for defining the meaningfulness of different effect sizes. He suggested that an effect size of 0.20 could be considered small, one of 0.50 could be considered medium, and one of 0.80 could be considered large. Now, as more and more researchers have presented their findings in the effect size metric, we know that Cohen's proposed standard were far too optimistic. Hill, Bloom, Black, and Lipskey (2008) recently reviewed the effects found by real researchers in real studies and found that effect sizes were shockingly small relative to the norms proposed by Cohen. In elementary schools, for instance, when effects were typically measured using standardized achievement tests, experimental treatment effects averaged only 0.07, far smaller than Cohen's proposed "small" effect of 0.20. When the content of the test was more narrowly defined, the average effect jumped to 0.23 and when it was even more specialized, it jumped to 0.44. As a result, education studies are more likely to find larger effects in middle schools and high schools, where tests are subject-specific and their content is more advanced.

These averages do not mean that larger effects are not possible, nor that we have not generated larger effects. They are *averages* across a wide range of educational interventions. Still, they give us a sense for the kind of outcomes we might expect, so the first thing we might ask about PD programs is how they compare with the effects of other kinds of education interventions. For this analysis, following Hill et al. (2008), I first sorted programs according to whether their outcome measures were broad or narrow. The majority of studies in my population were located in elementary schools and relied on broad standardized achievement tests. Only a handful used more narrowly focused outcome measures—one working with teachers of English language learners and a few others working on specific science topics.

Table 1 presents average effect sizes for studies using these two kinds of outcomes.[3] The top two rows of Table 1 provide two possible benchmark values against which to compare studies using broader outcome measures. First, we see the average effect size of all education studies examined by Hill et al. (2008), which was 0.07, and then we see the effect size of the "bootstrap" program described above, a program that simply pairs more effective teachers with less effective teachers.

After these rows in Table 1 are two rows showing the average effects of PD programs, first those that were evaluated with broad achievement tests in mathematics and language arts, and then those with specialized tests in the sciences or in English language learning. Finally, to help us understand the value of a second set of outcomes, the last line of Table 1 shows the average effect size Hill and others found when their studies used more narrow measures.

This comparison is not very encouraging. On average, our myriad expensive PD programs are almost indistinguishable from Papay et al.'s (2016) inexpensive bootstrap approach that encourages teachers to help one another. Yet most of these programs are far more expensive and time-consuming.

Still, the programs gathered here are quite various, and it behooves us to examine them further to see what else we might learn about the potential for PD to improve teaching practice. In the remainder of this chapter, I use patterns of program effects to address a variety of questions about how PD works and what we can expect from it.[4]

WHAT DO TEACHERS NEED TO LEARN?

The central premise underlying all PD is that there is something the researcher knows about teaching that teachers do not know. Over time, our hypotheses about what that is have shifted, and this shift largely reflects changes in how researchers themselves conceptualized the practice of teaching. So my first examination of PD literature sorts studies according to their hypotheses about what teachers need to learn. The most common hypotheses involve specific procedures, content knowledge, or strategies and insights.

Procedures

One stream of research focuses on what teachers are *doing*, often with little regard to *why* they did those things or to what their students were doing. This was the first approach we used to define the practices of teaching. A vocal advocate for this line of work, Nate Gage (1977) argued that the field needed a *scientific basis* for what had previously been thought of as "the art of teaching." To this end, researchers tried to partition teaching into a collection of discrete practices and then see which practices were correlated with student achievement gains. Once they became convinced that they had identified a set of such behaviors, they began devising PD programs to teach those behaviors to teachers.

TABLE 1
Overall Effects of Professional Development

Source of Study Effect Sizes	Average 1-Year Effect Sizes
Hill, Bloom, Black, and Lipskey's (2008) expected value for standardized tests	0.07
Papay, Taylor, Tyler, and Laski's (2016) bootstrap study	0.12 (1 study)
All elementary math or language professional development programs	0.10 (20 studies)
Topic-specific professional development programs	0.27 (4 studies)
Hill et al.'s expected value for narrower content tests	0.44

I found seven experimental studies that were designed to prescribe specific things teachers should do. Most focused on generic teaching practices such as questioning techniques or management techniques. One (Borman, Gamoran, & Bowdon, 2008) provided highly specified behavioral guidance on how to implement a new science curriculum. For instance, here is a passage from the teachers' manual describing a single fourth-grade unit (*Rot It Right: The Cycling of Matter and the Transfer of Energy. 4th Grade Science Immersion unit*, 2006):

- 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.
- 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?
- 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.
- Have students record and explain their predictions for each set of columns for later reference. (p. 21)
- . . .

Content Knowledge

The second stream of work focuses on teachers' content knowledge. Interest in content knowledge arose relatively early in our history of PD research, and it derived from a study of teaching *behaviors* (Good & Grouws, 1979). These authors tested a PD model that stipulated a sequence of lesson segments for mathematics lessons. The guideline suggested that teachers spend about 8 minutes reviewing concepts that had

been previously taught, then spend 20 minutes in "development," which involved introducing new concepts and questioning students to make sure they understood them, then move to a homework assignment, and so forth. But as the researchers watched teachers try to implement this relatively simple lesson framework, they came to see that teachers had great difficulty with the phase called "development." At first, the researchers tried to solve this problem by defining "development" more clearly, telling teachers that it was where they should attend to relationships between concepts and procedures, to students' confusions, and so forth. Ultimately, they concluded that teachers could not enact this step because they lacked sufficient *content knowledge*. A few years later, Lee Shulman (1986a, 1986b, 1987) actively sought to redirect the field away from behavioral depictions of teaching practice and toward a focus on teachers' knowledge.

This shift in attention from discrete teaching behaviors toward content knowledge is an example of researchers themselves learning about teaching. PD programs based on prescribed behaviors were part of an effort to be more scientific and objective, but they also represented a relatively naive conception of teaching, which is, after all, a "knowledge" profession.

Notice, though, that when I say that the field has shifted from behavioral admonitions to content knowledge, I do not mean that the former approach has entirely disappeared. In fact, the science program described just above is an example of a relatively recent addition to our PD oeuvre, but rather than teaching teachers science content, it teaches them the procedures they should use to teach that content.

I found five studies of PD that focused on teachers' content knowledge. Four addressed mathematical content (Garet et al., 2010; Garet et al., 2011; Garet et al., 2016; Jayanthi, Gersten, Taylor, Smolkowski, & Dimino, 2017; Niess, 2005); the fifth (Garet et al., 2008) focused on language arts. These programs tend to look a lot like conventional college classrooms, with teachers playing the role of students. There may be lectures, there may be question-and-answer sessions, there may be small-group discussions, and there may even be textbooks and homework.

Content knowledge continues to be a popular theme in discussions about what teachers need to know, but it is possible that our perception that teachers' content knowledge is seriously deficient could derive from attribution errors such as I experienced with Ms. Katlaski. This view is also relatively naive, in that it overlooks things like motivation, organization, representation, and so forth.

Strategies and Insights

The 1986 *Handbook of Research on Teaching* introduced yet another conception of teaching. It included a chapter about myriad decisions teachers make throughout their lessons (Clark & Peterson, 1986). This chapter depicted the practice of teaching as a process of continuous decision making as teachers interpreted and responded to events as they unfold in the classroom. It suggested that much of what we observed in classrooms was not planned behavior but rather spontaneous responses to unfolding events.

The realization that the behaviors we see might be contingent on circumstances led to yet a third approach to PD, one that focused on how to interpret events as they unfolded and how to respond to them strategically. Instead of prescribing specific teaching moves, these programs offered insights into how students make sense of their lessons and offered broad strategies for engaging and responding to their students. They often provide coaches or mentors who could visit teachers within their classrooms or to convene groups of teachers locally to share and compare their experiences.

I found 17 programs that focused on insights and strategies. They are quite various in their design, but a prominent theme has to do with gaining a deeper understanding of how students think and why they say or do what they do.

The earliest study in this group (Carpenter, Fennema, Peterson, Chiang, & Loef, 1989) focused on elementary school mathematics. The authors met with teachers in a small discussion group and showed teachers a series of videotaped interviews with students. The tapes were intended to show teachers how students made sense of mathematical relationships. As the group examined and discussed these videos, teachers gained better understanding of the content itself, but they also developed a better understanding about how their students thought about that content and how to interpret the things their students said. The program made no prescriptions about what teachers should actually *do.* That was largely up to them.

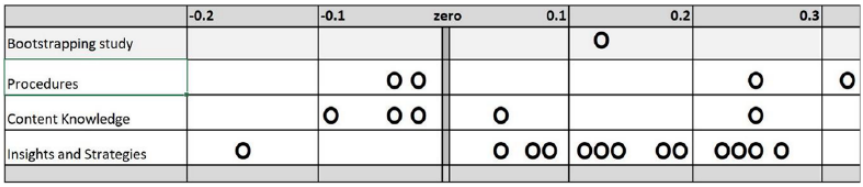
The proliferation of PD programs focusing on strategic teaching again reflects what we researchers ourselves have learned about teaching. As we have become more aware of the many contingencies involved in teaching, we have shifted our focus away from telling teachers what to do and toward deepening their understanding of their students so that they can make better in-the-moment judgments about how to respond to students.

PROGRAM EFFECTS

Do these different conceptions of teaching make a difference? The easiest way to compare these three approaches to PD is to focus on a subset of programs with common study designs. In this case, I compare 22 programs that all worked with teachers for a single academic year and used a general achievement test to measure student achievement. Figure 1 arrays the effects of these programs along a horizontal scale. Each program is characterized by a single circle. The figure does not include the science or English language learner studies, where test metrics can yield different effect sizes, or the studies that worked with teachers for only a portion of the year.

You can see that program effects are quite various across these studies, and that many of them appear to be less effective (though more expensive) than the bootstrap program, whose effect of 0.12 is shown in the top row. However, a larger fraction (2/3) of studies in the third group program yielded program effects larger than the 0.12 benchmark.[5]

FIGURE 1
Distribution of One-Year Program Effects by Conceptions of Teaching



The third conception also appears to be more widespread, with 13 research groups testing programs based on this vision of teaching. Thus we might benefit from a closer look at this approach to PD. Teachers are nearly always aware of multiple things continuously unfolding in their classrooms. They see that Ronald is confused, that Juanita is getting restless, that Mark is eager to show off what he has figured out, that someone has spilled something sticky on the floor near her desk, that the room is getting too hot because the janitor has not yet fixed the heater, and that the lunch bell will ring in 10 minutes. Regardless of what teachers plan to do during a given lesson, their in-the-moment actions are often responses to these in-the-moment observations. They want to suppress the show-off, calm down the fidgeter, help the confused student, open the window, and make sure the lesson reaches an appropriate closure before the bell rings. Then the mess can be cleaned up.

Strategic thinking is not merely about finding the best way to achieve the lesson goal; it is also about *seeing* things that might interfere with or facilitate the direction of the lesson, watching for signs of restlessness or confusion, inventing ways to avoid or capitalize on these moments, and generally being aware of what all the students are thinking and doing. Much of the strategically oriented PD had to do with interpreting students' comments and recognizing signs of confusion or disorientation that need to be addressed. I suspect that one reason why strategically oriented PD was more successful is that it helped teachers get better at *seeing signals* within their own classrooms.

These programs offer two thing that are often missing when programs focus on procedures or content knowledge. One is classroom artifacts. Many of these programs rely on videotapes of classroom events, examples of student work, interviews with children, or other artifacts that demonstrate to teachers the issues on which they want to focus. Thus, conversations about teaching are not about universal methods but about interpreting and responding to specific types of situations. Second, the people who provide this sort of PD tend to be people who have themselves spent a great deal of time in classrooms and are cognizant of all these nuances of classroom life. They themselves have an intimate familiarity with the complications of teaching. They are not merely telling teachers what to do or what to say; they are showing teachers what to *look for*. Which raises another question:

Can We "Package" Successful PD Programs?

An important reason for conducting research is to be able to identify effective practices so that others can adopt them. We want to be able to "package" effective programs and distribute them to a wider audience. But what if program effectiveness depends on the PD provider's own personal knowledge of classroom life, on his or her ability to spontaneously generate examples or to spontaneously notice things while visiting teachers' classrooms? If the quality of the message depends on the provider's intimate knowledge of classroom life, other providers, even when trained in the PD approach, might not be able to achieve the same outcomes.

Figure 2 presents a rough attempt to test this "intimate knowledge" hypothesis. It separates programs provided by their original developers from programs that were packaged by institutions. Figure 2 shows the same array of program effects as Figure 1, except in this case, packaged programs are marked with "X's" rather than with "O's."

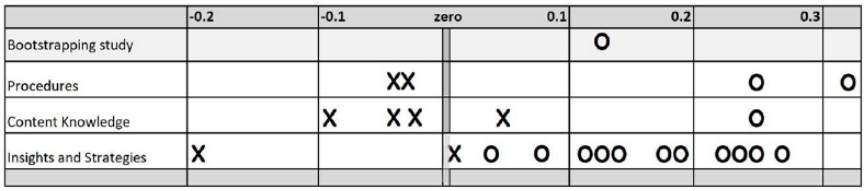
Notice that almost all the "X's" reside on the negative end of the distribution, while the "O's" are all on the positive side. This pattern creates an interesting dilemma, for it suggests that our large-scale studies, the kind of studies researchers value most, are not effective at raising student achievement. There may be many reasons for this disparity, of course, having to do with the logistical difficulties of orchestrating large-scale studies or with the kinds of study samples used, but since my purpose is to examine our assumptions about teacher learning, I want to focus here on the hypothesis that these differences derive from packaged PD.

What exactly do I mean by "packaged" PD programs? In the following paragraphs, I contrast pairs of programs within each row, one program that I consider to be locally developed with a counterpart that appears to be packaged. To the extent possible, I strive to pair programs whose content and goals are comparable. However, I have not actually tested these pairs of studies to see whether their differences are statistically significant.

Procedural Knowledge

In one of the first studies of PD ever conducted, a group of researchers (Anderson, Evertson, & Brophy, 1979) generated a list of procedures that had been shown to be related to student learning, converted these into a list of recommended practices, and accompanied each recommendation with very brief rationale. For instance, one said, "The introduction to the lesson should give an overview of what is to come in order to mentally prepare the students for the presentation." Another said, "It is also at the beginning of the lesson that new words and sounds should be presented to the children so that they can use them later when they are reading or answering questions." The PD itself was remarkably brief, consisted of a single 3-hour orientation, during which the principal investigator presented the list as a whole, discussed its use, and allowed teachers to ask questions. They then asked teachers to try to use these admonitions for the entire school year. The program had a yearlong effect of 0.24 on student achievement, the second-most effective procedural program shown in Figure 2.

FIGURE 2
Distribution of Program Effects for Packaged Versus Original Programs



Now for the contrast: A few years after that study was done, another group (Coladarci & Gage, 1984) took the same list of admonitions and *mailed* it to a group of teachers to see if they could get the same effect. I consider this mailed list of admonitions to be a packaged message in part because it is more impersonal but also because there was no opportunity to discuss or clarify any of the admonitions, help teachers envision the kind of situations where they would be applicable, or respond to any questions. This mailed-in program had an effect of –0.04, compared to the 0.24 from the original study.

Subject Matter Knowledge

In general, all of the programs providing content knowledge looked roughly like college courses: Meetings were held in classroom-like settings, PD providers gave lectures and demonstrations, engaged in question-and-answer sessions, and formed teachers into small groups to solve practice problems. Sometimes programs also provided local coaches who visited teachers in their classrooms. This classroom format is not surprising; it is the customary way subject matter has always been taught, and it fits our perception of subject matter knowledge as universal, residing outside of specific situations. Yet only one of these programs was effective.

How did this program differ from the others? It is the only one that relied on local faculty who knew the local community, knew the schools and teachers, and could gear their presentations to these audiences. Furthermore, they were not given their curricula but instead taught courses they had designed themselves. They also tailored the program for teachers by frequently modeling the teaching of their content and by sponsoring an online forum where teachers could discuss issues with one another throughout the academic year. These activities helped teachers translate the content into their own situations.

In contrast, the less effective programs provided a uniform curriculum to all localities and hired presenters to teach this prespecified curriculum. Nothing in the programs was tailored to the unique needs or interests of participating teachers, nor is it clear whether the hired presenters were allowed to modify their program in response to the unique needs of their audiences. Nor is it clear whether the presenters had the kind of intimate knowledge of classroom life that would enable them to

generate spontaneous teaching examples or story problems that would be meaningful to teachers.

Strategies

I use the term *strategic* to distinguish programs that are focused on interpreting events and adapting instruction to circumstances. In general, strategies are more flexible than procedures, more responsive to unique circumstances, and more responsive to differences among students.

One of the most effective of these programs (Gersten, Dimino, Jayanthi, Kim, & Santoro, 2010) introduced first-grade teachers to research findings regarding early reading instruction, and did this by helping teachers incorporate these findings into their local lesson plans. Teachers met in groups throughout the academic year to jointly plan their reading lessons, and throughout these meetings, they were regularly introduced to new research findings. Each planning meeting followed a four-step process: First, teachers would report what happened when they implemented their previously planned lessons. Then they would discuss their newest report on research findings. In this phase, the group facilitator focused their attention on the central concepts to make sure everyone understood them. In the third phase, they would review the publisher's recommended lesson and discuss its strengths and weaknesses. Finally, they would work together to design a lesson of their own that incorporated the research principle they had just read about. Thus, in this PD, even though the program gave teachers a standardized curriculum, it did so by embedding the content into the lesson planning process. Each planning group made sense of the findings in the context of its own classrooms and then directly applied the new knowledge into their next lessons.

There is another program that also taught teachers about findings from reading research, but instead of working with teachers and helping them incorporate the findings into their lesson plans, this program packaged the material and presented it to teachers through a series of daylong seminar sessions, each accompanied by a textbook. So both programs wanted teachers to get better at teaching language arts, and both aimed to do so by introducing teachers to research findings in that area. The first had an effect size of 0.23 and appears in the "strategy" row of Figure 2, while the second had an effect of 0.05 and appears in the "content knowledge" row of Figure 2.

By definition, strategic programs are less amenable to packaging. They aim to engage teachers in classroom-based problem-solving and to help them "see" their own classrooms differently, a goal that seems to require program faculty who have intimate familiarity with classroom life, so much so that they can help their teachers interpret their own experiences differently.

One program in this group has been working toward standardization for several years, and its progress might be instructive here. The Cognitively Guided Instruction program, or CGI, was initially designed and tested by a group of mathematics faculty and graduate students at the University of Wisconsin in 1989 (Carpenter et al., 1989). At that time it was a unique program, one of the first programs to move away from

direct instruction and toward strategic thinking. It had a modest effect of 0.13. But the authors, along with their colleagues and graduate students, continued to use CGI in their college courses and in local PD programs for many years and to expand its influence. After about 10 years, they developed a guide for workshop leaders (Fennema, Carpenter, Levi, Franke, & Empson, 1999). Then, after another 10 years had passed, the younger generation of CGI mathematics educators (Jacobs, Franke, Carpenter, Levi, & Battey, 2007) carried out a second experimental test of CGI and achieved a much higher effect of 0.26. Presumably, this improvement reflected a series of refinements over time as all the members of this group became more familiar with teachers and their needs.

After that, some members of this group decided to create a formal organization to provide PD and related services. Called the Teachers Development Group, this organization sought to further disseminate the concept of CGI by providing written materials and making PD available on a broader scale. In other words, they sought to *package* the CGI program. But large-scale expansion runs the risk of relying on inexperienced PD providers who may have neither the personal, situated understanding of the program that the founders had nor the intimate knowledge of how teachers responded to CGI.

Now we have yet a third test (Schoen, LaVenia, Tazaz, & Faraina, 2018) of CGI, this one based on the new packaged version of the program. This new packaged version of CGI yielded an average effect of zero.

These three tests of CGI represent three different levels of PD provider experience. In the first test, yielding an effect of 0.13, the providers had knowledge of student learning and had experience teaching teachers in their classes but had no experience providing a PD program. In the second, yielding an effect of 0.23, program staff had knowledge of how children learn as well as more experience providing PD. But in the third study, yielding an average effect of 0.0, the program had been packaged for large-scale distribution, and I suspect local providers lack the kind of intimate familiarity with classroom life that is needed to help teachers alter their perceptions of their own experiences.

The pairs of outcomes I share here, of course, could reflect nothing more than ordinary statistical variations. However, the *pattern* of differential program effectiveness, across over 20 independent studies, raises important questions about the reliability of program effectiveness and, ultimately, about value of our PD research if our findings cannot be reliably expanded or replicated. Even if my hypothesis about packaging is wrong, we still need to think more about how we define salient program features that should be part of any replication effort and whether program staff experiences are a necessary "feature" of the program.

HOW DOES NEW KNOWLEDGE "TRAVEL" FROM PD TO STUDENT LEARNING?

The three groupings I outlined above suggest that different programs have different tacit theories about the kind of change that is needed. Some PD providers believe

that if they teach a set of specific procedures to teachers, and teachers implement them, those specific behaviors will foster student learning. If we teach content knowledge, we are assuming that teachers will be more able, on their own, to teach that content. But we still know very little about how to actually foster these changes, or about how much time is needed to foster such change. In an effort to help teachers make these changes, many PD providers send mentors or coaches into the schools, people who visit teachers within their classrooms and help them "see" new things and try new things. Thus, we may think about PD as having a cascading sequence of influences that looks like this:

PD → Coaches → Teachers → Students

The modal PD program works with teachers throughout a single full academic year, implying that researchers expect teachers to be able to alter habits and routines relatively quickly, and adopt their new recommendations relatively quickly. But there is another timing problem inherent in this approach to PD: Researchers typically measure changes in student achievement *during that same academic year*. This schedule is popular in part because student achievement is typically measured by school districts at the end of each academic year. So a PD provider who comes in, say September, might consider last spring's school test as his pretest. Normally, we think of causes as preceding effects, so this schedule raises a variety of questions: How quickly do we expect teachers to alter their practice based on what they have just learned? Do we expect them to alter their methods the next day? Within a week or a month? On the other hand, if change is slow, and if teachers need time to alter their habits, can we expect to see the effect of that change on student achievement gains that are measured concurrent with the treatment itself?

These complications with PD research designs invite questions about the array of program effects, for virtually all of them could be underestimating program effects: Students' annual achievement gains are almost always the result of teaching events that occurred before the PD has had its full influence.

But there are also scenarios that would lead us to *overestimate* effectiveness: Suppose teachers privately dislike the approaches being taught but comply with them only to be polite or to get their coaches to leave them alone. If this occurred, we might see a gain during the program year, but the gain would reflect *compliance* rather than genuine learning and it would go away the following year.

These scheduling problems provide another example of an area in which we need to learn more about *how to learn* about teacher learning, how to design our studies, and how to map exposure to PD with changes in practice and, in turn, changes in student learning. Most important, we need to learn more about whether program effects are sustained over time, and whether they accumulate over time.

Do Program Effects Last Over Time?

The modal study design, which measures student learning concurrent with program implementation, is built on the tacit assumption that learning is immediate throughout the program year. But much of teaching practice is habitual, and teachers may need more time to generate new practices.

Furthermore, the role of time may vary across programs. Those offering procedures might be hoping to save teachers the problem of translation by providing precise behavioral guidance in the first place. Those offering content knowledge skirt the question of whether behaviors need to change. Those offering insights, on the other hand, depend heavily on teachers' own intentions to determine what gets changed. Teachers could reject a new idea altogether or, conversely, discover more situations where the new insight applies.

If we could follow changes in student learning across multiple years, we might see that different patterns of *teacher* learning yield different patterns of *student* learning as well.

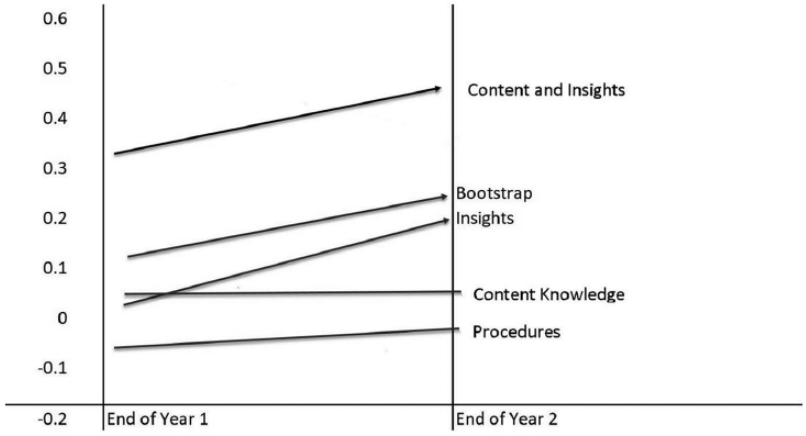
Only five studies followed teachers into a second year and measured their students' achievement during that next year. These studies provide some clues about what happens to new ideas over the long term. That is, assuming that teachers benefit from a program in the first place, is the effect sustained into a new year? Figure 3 plots changes in student achievement from the end of Year 1 to the end of Year 2. That is, the beginning of the line represents program effects at the end of one year. From that starting point, a horizontal line means program effects were sustained through the second year, and while downward trending lines might mean that teachers either forgot or purposefully abandoned the program's ideas, an upward trend might suggest that teachers not only sustained their knowledge but also continued to find more ways to incorporate new ideas into their practices so that students benefitted even more during the follow-up year.

There are only slight differences among these lines, suggesting that in general, teachers in all programs roughly sustained what they had learned during Year 1. Those programs that had negligible effects during Year 1 had almost the same negligible effects at the end of Year 2. Their lines are virtually horizontal.

Programs that had a moderate effect in Year 1, however, invite some interesting hypotheses about teacher learning. They imply that teachers continued to improve their effectiveness during the second year, even though their programs were no longer helping them. One hypothesis might be that teachers may need time to digest new ideas and to fully incorporate them into their practices. If so, the traditional 1-year study may not be sufficient to fully understand program effects. One of these delayed effects came from the bootstrap program. Below, I examine the other two.

The topmost line represents results from a program in science (Heller, Dahler, Wong, Shinohara, & Miratrix, 2012). The study addressed only a single unit (electricity) in the science curriculum and provided teachers with knowledge about electricity as well as deeper insights into how students learn about electricity. For the

FIGURE 3
Delayed Effects From Different Approaches to Professional Development



'insights" part, they tested three approaches: Some teachers examined their own students' work, others examined written cases of real teaching episodes, and still others examined their own experiences as learners. All three approaches had strong effects, and Figure 4 shows their average effectiveness.

The other program (Allen, Pianta, Gregory, Mikami, & Lun, 2011) consisted of ongoing consultations between teachers and mentors. Teachers videotaped sample lessons approximately every 2 weeks and sent their tapes to an online "teaching partner." Then the two of them would talk about the lesson. The nature of these conversations helps us understand the difference between prescriptions and insights. Instead of correcting teachers' behaviors, prescribing recommended practices, or evaluating what they saw on the video, these mentors used "prompts" to help teachers examine and think about specific events that had occurred. 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?" And 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 directly recommending any specific procedures or rules for teachers to follow, but there was a set of concepts the mentor wanted teachers to understand. Teaching partners posed questions that might help teachers think harder about their classroom experiences, about the relationship

between their own behaviors and the behaviors of their students, and about the enacted meaning of these concepts.

This kind of conversation, of course, requires that mentors themselves must be able to select revealing moments for examination, and must be able to pose provocative questions rather than recommend specific behaviors. If such a program wanted to expand, it would not be easy for them to hire more mentors, or even to define their selection criteria. As PD providers shift their programs away from procedures and knowledge and toward strategic thinking, they depend more and more on PD providers who themselves have enough depth of experience that they can recognize "teachable moments" within the PD process.

Do Program Effects Accumulate Over Time?

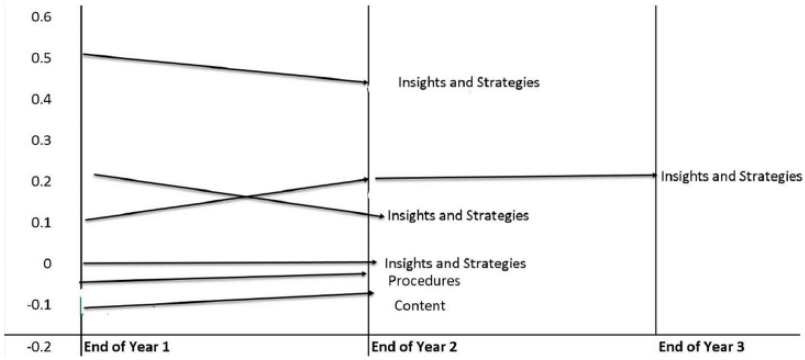
Although the bulk of programs spent a single academic year with teachers, there were a handful that continued to interact with teachers for a second year, and one remained for a third year. These programs may have added more content to their curriculum during their second and third years, or perhaps they used that time to reinforce their original ideas, making sure teachers would not fall back into bad habits from the past. In either case, these programs provide an opportunity for us to look at potential cumulative effects. That is, do additional years of engagement yield additional improvements in student learning?

Six programs spent more than one academic year with their teachers. Five spent a second year and the sixth spent both a second and a third year. Their results are shown graphically in Figure 4.

As with Figure 3, the beginning of each line reflects the program effects at the end of the first year of the program. Overall, this chart looks remarkably similar to Figure 3 in its distribution of Year 1 effects. Both charts include a couple of programs whose first-year effects were relatively small, and whose effects didn't change much during Year 2. Together, these charts suggest that programs with weak first-year effects failed to produce either delayed effects or cumulative effects later on.

The remaining slopes, those that started with greater Year 1 effects, show only slight changes during Year 2 and may reflect nothing more than random variations in outcomes. They certainly do not suggest that extended programs are adding a noticeable benefit. But a comparison of Figures 3 and 4 invites some interesting hypotheses about teacher learning. Figure 3 suggests that teachers might be continuing to grow during Year 2 as they find more ways to incorporate new ideas into their practices. That is, when programs spend one year with them, teachers then spend the next year further refining their new understandings and further improving their practice. But Figure 4 suggests that when programs spend a second year with them, the program may actually interfere with teachers' need to consolidate their new knowledge, so that second-year program effects tend to drift downward. While the data we have so far are sketchy, they do suggest that we need to design longer term studies if we are to gauge the full effects of our programs.

FIGURE 4
Cumulative Program Effects



WHERE TO NEXT?

The first study I described here (Anderson et al., 1979) was conducted almost 40 years ago, and I suspect it was the first experimental study of PD ever published in an educational journal. In the intervening years, education researchers have continued to pursue questions about what makes one teacher better than another, and about how we can provide guidance that would help teachers improve their practice. Many of our efforts have been naive in the sense that we thought teaching was much simpler than it has turned out to be.

I sorted these PD programs into three ways of thinking about how to improve teaching: one focusing on teaching behaviors, one on increasing content knowledge, and one on strategic thinking. The evidence we have now suggests that the third approach has had the greatest positive impact on teachers' effectiveness. Furthermore, there is some evidence that this approach enables teachers to *continue* to improve their own practice independently after the formal PD is finished. I suspect that the reason for this delayed success has to do with its emphasis on purpose, which in turn helps teachers function autonomously after the PD providers are gone.

I hope over time it will become customary for PD researchers to follow teachers for at least one full school year beyond the program's duration. As Huberman (1993) pointed out a long time ago, teachers are essentially tinkerers. They are accustomed to working in isolation, they depend heavily on their own personal innovations, and they depend on automated habits and routines. It makes sense, then, that they would need time to incorporate new ideas into their habits and routines. Though a few studies have followed teachers for a year beyond their treatment, the data shown here are too skimpy to yield any firm conclusion.

An important remaining problem has to do with replication. The most effective PD programs appear to be designed and carried out by people who have gained deep

personal knowledge of the intricacies of teaching. The patterns shown above suggest that their effectiveness is at least partially a function of this intimate knowledge. It is not clear whether or how PD providers can share this form of knowledge with other PD providers, thus raising questions about whether these programs can be expanded very much. We have reached a situation in which our knowledge about how to conduct productive PD is increasing but our ability to spread that knowledge is not. Meantime, teachers are being "treated" with ever-increasing volumes of packaged PD, at great expense to school districts and with almost no benefit for themselves or their students.

NOTES

[1] The present population of studies differs from the 2016 review as follows: (a) It excludes four studies whose samples included fewer than 20 teachers; (b) it removes one study that did not use random assignment and that I had mistakenly included earlier; (c) it includes four studies that followed teachers for less than a full academic year (my 2016 criteria required a minimum full academic year minimum); and (d) it adds six studies published since that review was completed.

[2] Randomization can be done in many ways. Researchers may assign individual teachers, whole school populations, or subgroups of teachers within schools. Sometimes they solicit volunteers first and then assign only volunteers to groups. The most common mistake in PD research is to solicit volunteers for their program, then seek out a group of *seemingly* comparable teachers for a comparison. This design overlooks the importance of motivation to learn as a factor in learning, and I rejected all of the studies based on matched groups.

[3] Readers are referred to my earlier article (Kennedy, 2016) in the *Review of Educational Research* for computational details.

[4] The following analysis is not intended to draw conclusions about relative program effectiveness but rather to use outcome patterns to generate hypotheses about teacher learning and about how PD fosters learning.

[5] I have not formally tested for differences among discrete program effects. Study sample sizes ranged from 20 to over 400 with more recent studies using larger samples.

STUDIES EXAMINED

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*, 19, 1034–1037.

Anderson, L. M., Evertson, C. M., & Brophy, J. E. (1979). An experimental study of effective teaching in first-grade reading groups. *Elementary School Journal*, *4*, 193–223.

Babinski, L., Amendum, S. J., Knotek, S. E., Sanche, M., & Malone, P. (2018). Improving young English learners' language and literacy skills through teacher professional development: A randomized controlled trial. *American Educational Research Journal*, *55*, 117–143.

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.

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

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.

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

Duffy, G. G., Roehler, L. R., Sivan, E., Rackliffe, G., Book, C., Meloth, M. S., . . . Bassiri, D. (1987). Effects of explaining the reasoning associated with using reading strategies. *Reading Research Quarterly*, *22*, 347–368.

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/20084031

Garet, M. S., Helpen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., . . . Yang, R. (2016). *Focusing on mathematical content knowledge: The impact of content-intensive teacher professional development*. Washington, DC: U.S. Department of Education.

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

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.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study*. Washington, DC: National Center for Education Evaluation. Retrieved from https://files.eric.ed.gov/fulltext/ED565837.pdf

Good, T. L., & Grouws, D. A. (1979). The Missouri Mathematics Effectiveness Project: An experimental study in fourth grade classrooms. *Journal of Educational Psychology*, *71*, 355–362.

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.

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.

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.

Jayanthi, M., Gersten, R., Taylor, M. J., Smolkowski, K., & Dimino, J. (2017). *Impact of the developing mathematical ideas professional development program on grade 4 students' and teachers' understanding of fractions*. Washington, DC: National Center for Educational Evaluation and Regional Assistance.

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.

McMeeking, L. B. S., Orsi, R., & Cobb, R. B. (2012). Effects of a teacher professional development program on the mathematics achievement of middle school students. *Journal for Research in Mathematics Education*, *43*, 159–181.

Myers, C. V., Molefe, A., Brandt, W. C., Zhu, B., & Dhillon, S. (2016). Impact results of the eMints professional development validation study. *Educational Evaluation and Policy Analysis*, *38* 455–476.

Niess, M. (2005). Oregon ESEA Title IIB MSP: Central Oregon Consortium. Corvallis: Department of Science and Mathematics Education, Oregon State University.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data*. Cambridge, MA: National Bureau of Economic Research.

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.

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

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

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

Schoen, R. C., LaVenia, M., Tazaz, A. M., & Faraina, K. (2018). *Replicating the CGI experiment in diverse environments: Effects of Year 1 on student achievement*. Tallahassee: Florida State University Learning Systems Institute.

Supovitz, J. (2013, April). *The linking study: An experiment to strengthen teachers' engagement with data on teaching and learning*. Paper presented at the American Educational Research Association conference, San Francisco, CA. Retrieved from https://files.eric.ed.gov/fulltext/ED547667.pdf

REFERENCES

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*, 19, 1034–1037.

Anderson, L. M., Evertson, C. M., & Brophy, J. E. (1979). An experimental study of effective teaching in first-grade reading groups. *Elementary School Journal*, *4*, 193–223.

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.

Bredeson, P. V. (2001). Negotiated learning: Union contracts and teacher professional development. *Education Policy Analysis Archives, 9*. Retrieved from https://epaa.asu.edu/ojs/article/viewFile/355/481

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.

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, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

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

Fennema, E., Carpenter, T., Levi, L., Franke, M. L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction: A guide for workshop leaders*. Portsmouth, NH: Heinemann.

Gage, N. L. (1977). *The scientific basis of the art of teaching*. New York, NY: Teachers College Press.

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/20084031

Garet, M. S., Helpen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., . . . Yang, R. (2016). *Focusing on mathematical content knowledge: The impact of content-intensive teacher professional development*. Washington, DC: U.S. Department of Education.

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.

Good, T. L., & Grouws, D. A. (1979). The Missouri Mathematics Effectiveness Project: An experimental study in fourth grade classrooms. *Journal of Educational Psychology*, *71*, 355–362.

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.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipskey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Huberman, M. (1993). The model of the independent artisan in teachers' professional relations. In J. W. Little, & M. W. McLaughlin (Eds.), *Teachers' work: Individuals, colleagues, context* (pp. 11–50). New York, NY: Teachers College Press.

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.

Jayanthi, M., Gersten, R., Taylor, M. J., Smolkowski, K., & Dimino, J. (2017). *Impact of the Developing Mathematical Ideas professional development program on grade 4 students' and teachers' understanding of fractions*. Washington, DC: National Center for Educational Evaluation and Regional Assistance.

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

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

Kennedy, M. M. (2010a). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*, 591–598.

Kennedy, M. M. (2010b). *Teacher assessment and the quest for teacher quality: A handbook*. San Francisco, CA: Jossey Bass.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*, 945–980.

Lortie, D. C. (1975). *Schoolteacher: A sociological study*. Chicago, IL: University of Chicago Press.

McDaniels, G. L. (1975). The evaluation of follow through. *Educational Researcher*, *4*(11), 7–11.

Niess, M. (2005). Oregon ESEA Title IIB MSP: Central Oregon Consortium. Corvallis: Department of Science & Mathematics Education, Oregon State University.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data*. Cambridge, MA: National Bureau of Economic Research.

Rot it right: The cycling of matter and the transfer of energy. 4th grade science immersion unit. (2006). Los Angeles, CA: SCALE. Retrieved from http://fastplants.org/pdf/scale/rotitright2006.pdf

Schoen, R. C., LaVenia, M., Tazaz, A. M., & Faraina, K. (2018). *Replicating the CGI experiment in diverse environments: Effects of Year 1 on student achievement*. Tallahassee: Florida State University Learning Systems Institute.

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

Shulman, L. S. (1986b). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 3–36). New York, NY: Macmillan.

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

TNTP. (2015). *The Mirage: Confronting the hard truth about our quest for teacher development*. New York, NY: Author. Retrieved from https://tntp.org/publications/view/the-mirage-confronting-the-truth-about-our-quest-for-teacher-development