Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures

Laura M. Desimone

The author suggests that we apply recent research knowledge to improve our conceptualization, measures, and methodology for studying the effects of teachers' professional development on teachers and students. She makes the case that there is a research consensus to support the use of a set of core features and a common conceptual framework in professional development impact studies. She urges us to move away from automatic biases either for or against observation, interviews, or surveys in such studies. She argues that the use of a common conceptual framework would elevate the quality of professional development studies and subsequently the general understanding of how best to shape and implement teacher learning opportunities for the maximum benefit of both teachers and students.

Keywords: causal inference; instructional practices; professional development; research methodology; survey research; teacher quality

> Before that time [the invasion of culture] the Romans were satisfied with the practice of virtue; they were undone when they began to study it.
>
> —Jean-Jacques Rousseau,
> *The Moral Effects of the Arts and Sciences*

This article offers ideas to improve the quality of inquiry into teacher learning, one of the most critical targets of education reform. Research increasingly has identified the continuing development and learning of teachers as one of the keys to improving the quality of U.S. schools (Borko & Putnam, 1995; Carnegie Forum on Education and the Economy, 1986; Darling-Hammond, 1993; The Holmes Group, 1986; National Commission on Teaching and America's Future, 1997; Talbert & McLaughlin, 1993; Thompson & Zeuli, 1999). It is also one of the critical mediators in the effectiveness of policy for teachers and teaching practice (Desimone, Smith, & Frisvold, 2007; T. M. Smith, Desimone, & Ueno, 2005) and in improving student achievement (Desimone, Smith, Hayes, & Frisvold, 2005). Many reforms rely on teacher learning and improved instruction to increase student learning;[1] in fact, education reform is often synonymous with teachers' professional development (Sykes, 1996). Substantial resources are spent on professional development at the local, state, and federal levels; for example, in 2004–2005, the federal government spent about $1.5 billion on professional development for teachers (Birman et al., 2007). Thus, understanding what makes professional development effective is critical to understanding the success or failure of many education reforms.

For decades, studies of professional development consisted mainly of documenting teacher satisfaction, attitude change, or commitment to innovation rather than its results or the processes by which it worked (Frechtling, Sharp, Carey, & Vaden-Kiernan, 1995; Guskey, 2000). In the past decade the field has acknowledged a need for more empirically valid methods of studying professional development. My arguments in this article are a response to this need. I discuss several key issues relevant to raising the quality of studies that assess how effectively professional development improves teaching practice and increases student achievement.

Specifically, I contend that the myriad of experiences that count as teacher learning pose a challenge for measuring professional development in causal studies, and I propose that measuring the core features of teachers' learning experiences is a way to address this challenge. I argue that there is an empirical research base to support the identification of a core set of features of effective professional development and a core conceptual framework for studying the effects of professional development; I describe what the core features and conceptual framework are, and cite research and policy documents to support their importance. I further argue that we should use the common conceptual framework, which includes the core features of effective professional development, as a base for effectiveness studies of professional development. I then suggest that one reason why a research consensus pointing to a common conceptual framework is not more obvious may be misconceptions about trade-offs of different methods used to study professional development's impacts. I hypothesize that these misconceptions might in part be due to "folklore" fostered by studies conducted in the 1960s and 1970s that attempted to validate measures of teachers' instruction. I then evaluate those early studies on the basis of current standards of evidence and conclude that they do not provide useful results. I argue that the research consensus is warranted, given current evidentiary standards, and that studies of professional development's effectiveness

Educational Researcher, Vol. 38, No. 3, pp. 181–199
DOI: 10.3102/0013189X08331140

should use the core features and conceptual framework I describe herein. Finally, in the context of these arguments, I outline areas for future work in professional development research.

Pressing Questions

> Human reason has this peculiar fate that in one species of its knowledge it is burdened by questions which, as prescribed by the very nature of reason itself, it is not able to ignore, but which, as transcending all its powers, it is also not able to answer.
>
> —Immanuel Kant,
> *Critique of Pure Reason*

Given the critical role of professional development in school improvement efforts, I pose the following research question: *How can we best measure professional development, and its effects on teachers and students, toward the end of improving professional development programs and policies to foster better instruction and student achievement?* To address this broad question, I examine three relevant subquestions:

1. What counts as professional development?
2. What purposes could a core conceptual framework serve, and what such framework is supported by the research?
3. What are the implications for modes of inquiry in causal studies of teacher learning?

What Counts as Professional Development?

Teachers experience a vast range of activities and interactions that may increase their knowledge and skills and improve their teaching practice, as well as contribute to their personal, social, and emotional growth as teachers. These experiences can range from formal, structured topic-specific seminars given on in-service days, to everyday, informal "hallway" discussions with other teachers about instruction techniques, embedded in teachers' everyday work lives.

Naturalistic and descriptive studies using ethnographic or in-depth case-study methods often allow the examination of nearly all learning experiences that a teacher has during a particular study period (e.g., Denzin & Lincoln, 2002; Merriam, 1988; Miles & Huberman, 1994; Spindler, 2000; Yin & Campbell, 2003). In contrast, studies that ask questions about trends, associations, or impacts require us to make a priori decisions that identify the teacher learning experiences on which we wish to collect data. How might we sensibly identify discrete learning experiences for such studies?

Defining Professional Development

The literature casts a wide net for what might be included as professional development, described by Little (1987) as "any activity that is intended partly or primarily to prepare paid staff members for improved performance in present or future roles in the school districts" (p. 491). Moving beyond discrete activities such as workshops, local and national conferences, college courses, special institutes, and centers (Little, 1993) are the newer, more complex and broad-based views on how to conceptualize teachers' professional development that have begun to emerge over the past decade. Situated and cognitive views of learning as interactive and social (Greeno, 1997; Greeno, Collins, & Resnick, 1996), based in discourse and community practice (e.g., Anderson, Reder, & Simon, 1996, 1997; Cobb, 1994; Greeno, 1997; Lave & Wenger, 1991), have been applied to teachers (Putnam & Borko, 2000). This is consistent with the idea that formal or informal learning communities among teachers can act as powerful mechanisms for teacher growth and development (e.g., Little, 1999, 2002; McLaughlin & Talbert, 1993; Stein, Smith, & Silver, 1999).

These newer conceptualizations imply challenges to the measurement of professional development at both the individual and community levels, typified by Paul Cobb's (1994) statement that "learning should be viewed as both a process of active individual construction and a process of enculturation into the . . . practices of wider society" (p. 13). Similarly, Hilda Borko's (2004) description of the myriad of contexts for teacher learning makes obvious the challenges of identifying and measuring teacher learning:

> For teachers, learning occurs in many different aspects of practice, including their classrooms, their school communities, and professional development courses or workshops. It can occur in a brief hallway conversation with a colleague, or after school when counseling a troubled child. To understand teacher learning, we must study it within these multiple contexts, taking into account both the individual teacher-learners and the social systems in which they are participants. (p. 4)

This type of embedded professional development, directly related to the work of teaching, can take the form of coteaching, mentoring, reflecting on actual lessons (Schifter & Fosnot, 1993), or group discussions surrounding selected authentic artifacts from practice such as student work or instructional tasks (Ball & Cohen, 1999; Gearhart & Wolf, 1994). Also, activities can come in the form of a book club (Grossman, Wineburg, & Woolworth, 2001) or a teacher network or study group (Greenleaf, Schoenbach, Cziko, & Mueller, 2001). Even curriculum materials are themselves a potential source of professional development when they are designed to be "educative" (Ball & Cohen, 1996; Loucks-Horsley, Hewson, Love, & Stiles, 1998; Remillard, 2005).

Furthermore, some of the most powerful teacher learning experiences can occur in a teacher's own classroom, through self- or observer examination of the teacher's practice (Putnam & Borko, 2000). Still another dimension of teachers' professional development is their individual activities, such as engagement in educative online venues (Ingvarson, Meiers, & Beavis, 2005) and their own inquiry/action research (Guskey, 2000).

Another kind of teacher activity that falls under the professional development umbrella is involvement in a development or improvement process (Guskey, 2000; Little, 1993)—for example, designing or choosing new curricula or textbooks or assisting with the school improvement plan. Guskey points out that learning opportunities for teachers occur every time a lesson is taught, an assessment is administered, a curriculum is reviewed, or a professional journal or magazine is read. These examples illustrate the dynamic nature of professional development as ongoing, continuous, and embedded in teachers' daily lives (Lieberman, 1995; Loucks-Horsley et al., 1987).

Given this array of complex, interrelated learning opportunities, how can we distinguish learning activities from each other in

studies designed to describe trends, associations, or impacts of professional learning on knowledge, instruction, and student achievement? The next section suggests that a partial solution to this challenge is to focus on the critical features of teachers' learning experiences rather than on their structure.

Measuring Professional Development in Impact Studies: A Critical Features Approach

One way of translating the complex, interactive, formal, and informal nature of teacher learning opportunities into manageable, measurable phenomena is to focus measurement on the *critical features* of the activity (Desimone, Porter, Garet, Yoon, & Birman, 2002; Garet, Porter, Desimone, Birman, & Yoon, 2001)—those characteristics of an activity that make it effective for increasing teacher learning and changing practice, and ultimately for improving student learning—rather than on the *type* of activity (e.g., workshop or study group).

A study of a national probability sample of teachers showed that the features of professional development were what mattered for relationships with changes in knowledge and skills and classroom practice. The effects of the structure of the learning opportunity on teacher change—for example, whether it was a workshop or study group—were fully explained by the features of the activity (Desimone, Porter, et al., 2002; Garet et al., 2001). This finding suggests the potential usefulness of focusing on measuring not the structure of the activity but the features of professional development that have been shown to be related to the outcomes we care about.

Such an approach would require a consensus on the core features. Determining whether there is an established consensus on the core features of high-quality professional development is not an exact science. Here I argue that (a) there is enough empirical evidence to suggest that there is in fact a consensus on a core set of features, and (b) given this consensus, these core features should be included in studies of the effectiveness of professional development, to allow studies to build on each other and refine and expand our knowledge base.

There is no clear guidance indicating the thresholds required to achieve "consensus." Part of the challenge in determining whether a research-based consensus exists is to distinguish ideas grounded in empirical study from those grounded in conventional wisdom and those based on conceptual/theoretical ideas (Ball, 1996). A clear delineation of such sources is often not possible in research on teacher learning, as the characteristics identified as effective are usually a mix derived from all three sources (e.g., Elmore, 2002; Little, 1993; Loucks-Horsley, Love, Stiles, Mundry, & Hewson, 2003; Putnam & Borko, 1997; Wilson & Berne, 1999).

Still another issue is what counts as causal evidence, where the literature includes study designs on the continuum from intensive ethnographic studies of a couple of teachers, to national correlational studies, to randomized field trials. For example, in recent work Penuel, Fishman, Yamaguchi, and Gallagher (2007) cite case-study research to show that we have already identified what might contribute to high-quality professional development, but indicate that before the Garet et al. (2001) study, we had no empirical evidence of the *relative* value of specific features of professional development. Similarly, 6 years ago Elmore (2002) said that "theorists and researchers of teacher education and education reform [had] developed a preliminary consensus, with some empirical support, about what constitutes effective teacher professional development for promoting standards-based instruction in K–12 math" (p. 144). We have no formal guidance about what it would take to move from a preliminary to an established consensus. And there is also debate about whether showing effects on teaching practice is enough to count a characteristic as effective, or whether only links to improved student achievement warrant the "effectiveness" label. We do not have sufficient evidence to indicate which features of professional development are effective for eliciting improvements in student learning (Wayne, Yoon, Zhu, Cronen, & Garet, 2008); this is all the more reason we need to systematically include features in studies of impacts on student achievement.

My argument that we do have a research consensus on at least five core features should be interpreted in the context of these ideas about how and when a research consensus is established. In the next section I highlight the case-study work of the past several decades that supports the core features, as well as the more recent correlational, quasi-experimental, and experimental work. Given the number, quality, and diversity of studies that provide support for the features, I conclude that we have reached a consensus that these core features play an important role in determining the effectiveness of professional development, that they are the "features of PD worth testing" (Wayne et al., 2008, p. 472), and that, as such, they should be included in impact studies. Their systematic inclusion in effectiveness studies will allow us to take the next step to understanding the relative importance of the features for improving student achievement in different contexts (see Wayne et al., 2008).

What Purposes Could a Core Conceptual Framework Serve, and What Such Framework Is Supported by the Research?

Working from the notion that teacher learning experiences come in a multitude of formal and informal, embedded and discrete activities, are some key features of these experiences essential to measure when we study teachers' professional development? Providing a thorough review of all that we know about what makes professional development effective is beyond the scope of this article. However, there is a research consensus on the main features of professional development that have been associated with changes in knowledge, practice, and, to a lesser extent, student achievement. These critical features form the basis of the framework I propose for studying the effectiveness of professional development. I describe them below.

The Promise of a Core Conceptual Framework for Studying Professional Development: Critical Features of Professional Development

Recent research reflects a consensus about at least some of the characteristics of professional development that are critical to increasing teacher knowledge and skills and improving their practice, and which hold promise for increasing student achievement (Hawley & Valli, 1999; Kennedy, 1998; Wilson & Berne, 1999): (a) content focus, (b) active learning, (c) coherence, (d) duration, and (e) collective participation. Very recent studies are already including this set of

core features as critical components of effective professional development (e.g., Jeanpierre, Oberhauser, & Freeman, 2005; C. Johnson, Kahle, & Fargo, 2007; Penuel et al., 2007).

Content focus. The content focus of teacher learning may be the most influential feature. A compilation of evidence in the past decade points to the link between activities that focus on subject matter content and how students learn that content with increases in teacher knowledge and skills, improvements in practice, and, to a more limited extent, increases in student achievement. This evidence comes from case-study data (e.g., Cohen, 1990), correlational analyses conducted with nationally representative teacher data (e.g., Garet et al., 2001; T. M. Smith et al., 2007), quasi-experiments (Banilower, Heck, & Weiss, 2005), longitudinal studies of teachers (e.g., Cohen & Hill, 2001; Desimone, Porter, et al., 2002), meta-analyses (e.g., Kennedy, 1998), and experimental designs (e.g., Carpenter, Fennema, Peterson, Chiang, & Loef, 1989).

Active learning. Opportunities for teachers to engage in active learning are also related to the effectiveness of professional development (Garet et al., 2001; Loucks-Horsley et al., 1998). Active learning, as opposed to passive learning typically characterized by listening to a lecture, can take a number of forms, including observing expert teachers or being observed, followed by interactive feedback and discussion; reviewing student work in the topic areas being covered; and leading discussions (Banilower & Shimkus, 2004; Borko, 2004; Carey & Frechtling, 1997; Darling-Hammond, 1997; Lieberman, 1996).

Coherence. The third core feature emphasized in the literature is coherence, the extent to which teacher learning is consistent with teachers' knowledge and beliefs (Consortium for Policy Research in Education, 1998; Elmore & Burney, 1997). The consistency of school, district, and state reforms and policies with what is taught in professional development is another important aspect of coherence (Elmore & Burney, 1997; Firestone, Mangin, Martinez, & Polovsky, 2005; Fullan, 1993; Guskey, 1994; Little, 1982; Penuel et al., 2007; Rosenholtz, 1991).

Duration. Research shows that intellectual and pedagogical change requires professional development activities to be of sufficient duration, including both span of time over which the activity is spread (e.g., one day or one semester) and the number of hours spent in the activity (Cohen & Hill, 2001; Fullan, 1993; Guskey, 1994; Supovitz & Turner, 2000). Research has not indicated an exact "tipping point" for duration but shows support for activities that are spread over a semester (or intense summer institutes with follow-up during the semester) and include 20 hours or more of contact time.

Collective participation. Another critical feature is collective participation. This feature can be accomplished through participation of teachers from the same school, grade, or department. Such arrangements set up potential interaction and discourse, which can be a powerful form of teacher learning (Banilower & Shimkus, 2004; Borko, 2004; Desimone, 2003; Fullan, 1991; Guskey, 1994; Little, 1993; Loucks-Horsley et al., 1998; Rosenholtz, 1989).

Policy Reflects the Consensus

Education policy documents within the past several years are beginning to reflect this research consensus on critical features of professional development. The No Child Left Behind Act of 2001 describes "high-quality" professional development as activities that "improve and increase teachers' knowledge of the academic subjects the teachers teach" (content focus) and that "are . . . sustained [and] intensive" (duration) and "are aligned with and directly related to state academic content standards, student academic achievement standards, and assessments" (coherence).

Similarly, the Teaching Commission (2004) report *Teaching at Risk: A Call to Action* emphasizes coherence (alignment) and collective participation (collaboration):

> Professional development should be *aligned* [italics added] with state and district goals and standards for student learning . . . and should also involve opportunities for *collaboration* [italics added] so that teachers can learn from each other. (p. 49)

Despite the evidence supporting the five core features, there is no core set of characteristics that researchers regularly measure in empirical studies of professional development. Sharing a conceptual framework that defines important features of teacher learning experiences has the potential to move the field forward in terms of building a consistent knowledge base. What might such a conceptual framework look like, and what advantages might it offer us?

Toward a Conceptual Framework for Studying Teachers' Professional Development

There are at least two central components to a conceptual framework for studying teachers' professional development. One is recognizing a set of critical features that define effective professional development, as discussed above. The second is establishing an operational theory of how professional development works to influence teacher and student outcomes. Such a theory would identify the key inputs and intermediate and final outcomes that characterize the effects of professional development. It would also identify the variables that mediate (explain) and moderate (interact to influence) professional development's effects.

I propose a basic model, shown in Figure 1, and recommend its use in all empirical causal studies of professional development. The model represents interactive, nonrecursive relationships between the *critical features* of professional development, teacher knowledge and beliefs, classroom practice, and student outcomes. As reflected in the figure, a core theory of action for professional development would likely follow these steps:

1. Teachers experience effective professional development.
2. The professional development increases teachers' knowledge and skills and/or changes their attitudes and beliefs.
3. Teachers use their new knowledge and skills, attitudes, and beliefs to improve the content of their instruction or their approach to pedagogy, or both.
4. The instructional changes foster increased student learning.[2]

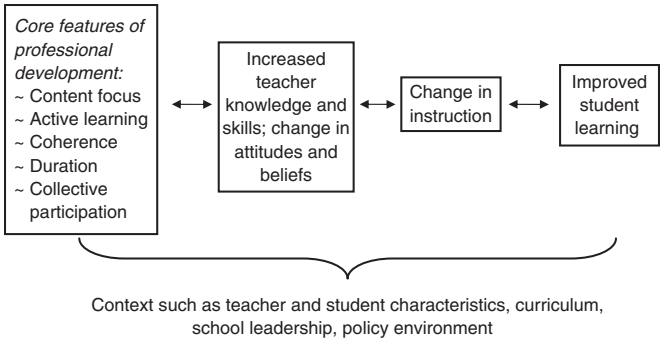

FIGURE 1. *Proposed core conceptual framework for studying the effects of professional development on teachers and students.*

This model allows testing both a theory of teacher change (e.g., that professional development alters teacher knowledge, beliefs, or practice) and a theory of instruction (e.g., that changed practice influences student achievement), both of which are necessary to complete our understanding of how professional development works (Wayne et al., 2008).

The importance of each element in my "path model" is reflected in the literature: links between teacher knowledge, practice, and student achievement (Hill, Ball, & Schilling, 2008; Phelps & Schilling, 2004; Snow, Burns, & Griffin, 1998; Wilson & Berne, 1999); instruction and student achievement (e.g., Hamilton et al., 2003; Mayer, 1998; Stein & Lane, 1996; Supovitz, 2001; Von Secker, 2002; Wenglinsky, 2002); professional development and teachers' practice (Fishman, Marx, Best, & Tal, 2003; Heck, Banilower, Weiss, & Rosenberg, 2008; Jeanpierre et al., 2005; Supovitz & Turner, 2000); and professional development and student achievement (Angrist & Lavy, 2001; Bressoux, 1996; Cohen & Hill, 2000, 2001; Jacob & Lefgren, 2004; O. Lee, Deaktor, Enders, & Lambert, 2008; Wiley & Yoon, 1995). A handful of studies have addressed links in all four areas illustrated in the figure—professional development, content knowledge, instruction, and student achievement (Carpenter et al., 1989; Franke, Carpenter, & Levi, 2001; Saxe, Gearhart, & Nasir, 2001).

Although empirical studies that include all elements are rare, the basic components are nearly universal in theoretical notions of the trajectories of teacher learning (e.g., Borko, 2004; Ingvarson et al., 2005), with variations that include an emphasis on context (Borko, 2004), changing the order to reflect teacher change in beliefs as a function of improved student achievement (Guskey, 2002), and acknowledgment of multiple pathways and individuality of teacher growth (Clarke & Hollingsworth, 2002). My notion of nonrecursive, interactive pathways does not prevent differential emphases on either the basic components (professional development, knowledge, practice, and student achievement) or the addition of moderating and mediating elements, such as teacher identity, beliefs, and perceptions.

The model operates with context as an important mediator and moderator. An examination of the literature identifies a strong consensus on several key mediating and moderating influences, reminiscent of Schwab's commonplaces of student, teacher, subject matter, and milieu (Schwab, 1973): (a) student characteristics such as achievement and disadvantage (Darling-Hammond & Sykes, 1999); (b) individual teacher characteristics, such as experience, knowledge, beliefs, and attitudes (Borko & Putnam, 1996; Carpenter, Franke, & Levi, 1998; Cohen & Ball, 1990; Feiman-Nemser, 1985; Grossman, 1990; Porter, 1989; Richardson, 1996); (c) contextual factors at the classroom, school, and district levels (Darling-Hammond & McLaughlin, 1999; Firestone, 1996; Grossman et al., 2001; Little, 2002; Newmann & Associates, 1996; Schultz, Jones-Walker, Chikkatur, 2008; Stein, Silver, & Smith, 1998; Thomas, Wineburg, Grossman, Myhre, & Woolworth, 1998); and (d) policy conditions at multiple levels (Desimone, 2002; Desimone, Birman, Porter, Garet, & Yoon, 2003; Desimone, Garet, Birman, Porter, & Yoon, 2002; Porter, 1994; Spillane, 2004).

Synthesizing Conceptual Frameworks to Form a Foundational Framework

As recently as last year, Borko, Jacobs, Eiteljorg, and Pittman (2008) indicated that there is little agreement about how to assess the quality of professional development. In response, I am proposing that our knowledge base has advanced to the point that we are justified in using a core framework to assess the effectiveness of professional development.

Several authors offer conceptual frameworks for studying teacher learning. My thesis is that, although we use different language and examine teacher learning from different perspectives and depths, there is a foundational conception present in most studies, whether they are conceptual, empirical, or both, which points to the common framework that I am proposing.

Borko's (2004) conception includes program, facilitators, teachers, and context—which map to the quality of professional development, as she indicates that an important part of what matters about the program and facilitators are their quality—and teacher characteristics and knowledge. She targets the core features I previously described, although her terminology is different. For example, she emphasizes a focus on subject matter and how students learn that subject matter, "engaging teachers as learners" (active learning), and strong professional learning communities (collective participation).

Peressini, Borko, Romagnano, Knuth, and Willis (2004) offer a conceptual framework that infers complex reflexive relationships between teaching practices and teachers' developing knowledge and beliefs about math, math-specific pedagogy, and professional identity. This model includes the knowledge, practice, student learning components, and an emphasis on subject-specific content, and is consistent with the notion of interactive paths. Wilson and Berne (1999) suggest three features of effective professional development: "communities of learning," teachers playing an active role, and "critical colleagueship" where trust and critique are present. In my view these three features overlap with notions of collective participation that result in opportunities for teachers to share and discuss, and active learning opportunities where teachers lead professional development. Thus, again we are talking about the same elements but using different language. Varied terminology and slight differences in construct definitions may be useful from a scholarly perspective to offer unique and nuanced insights, but it is arguably less useful to educators trying to make sense of the professional development literature.

In synthesizing the frameworks, it is clear there are several potentially important components not included in the base model that I propose, as they have not yet been subject to much impact research. These include professional identity (Peressini et al., 2004); the use of student work in professional learning (Borko, 2004) for reasons beyond its function of providing a content focus; the role of the principal in shaping learning opportunities and providing resources, time, encouragement, and monitoring (Banilower et al., 2005; Blasé & Blasé, 2000; Elmore & Burney, 1996; Guskey & Sparks, 1991; Nir & Bogler, 2008; Scribner, 1999); the role of curriculum materials and implementation (Banilower et al., 2005; Banilower & Shimkus, 2004; Penuel et al., 2007; Remillard, 2005); high expectations of professional development facilitators (Jeanpierre et al., 2005); and teacher reflection (Fishman et al., 2003). Future work may establish one or more of these as warranting inclusion in a foundational, common model of teacher learning, but as yet these features are not backed by as many substantial conceptual and empirical studies as are the core features (e.g., content focus, active learning) suggested in Figure 1.

Justification for a Core Conceptual Framework

A core conceptual framework is warranted for a number of reasons. First, as discussed in the preceding section, the literature seems to agree on what the basic elements of that framework should be.

Second, having a core set of characteristics that we know are related to effective professional development, and measuring them every time we study professional development, would help move the field forward. This is not to say that each study should be prevented from having its own unique characteristics or would be constrained to measuring only the features in a core framework. But the research consensus is strong enough to warrant the inclusion of a firm set of features that have been shown repeatedly, in case-study as well as large-scale and experimental research, to be related to teacher improvement and tentatively to student achievement. Even so, choosing what to measure and how to measure it requires interpretation. A shared conceptual framework could help guide measurement choices and help establish consistency that would contribute to building a knowledge base. Using a shared conceptual framework as a basis for developing measures of professional development would contribute to our building a consistent set of data over time on critical aspects of teachers' learning experiences.

Third, we need such a foundation to answer the types of questions called for in the field—such as indicating the relative impacts of various professional development interventions, determining how much and what types of professional development are needed, and figuring out the relative importance of particular features in various contexts (Banilower et al., 2005; Wayne et al., 2008).

Fourth, a shared conceptual framework could steer us toward using an appropriate timeline. As Loucks-Horsley et al. (1998) say, "It is foolhardy to either expect or focus on measuring student learning when teachers have just begun to learn and experiment with new ideas and strategies" (p. 222). Using a framework that suggests a sequence of events—from learning activities to changes in knowledge, beliefs, and attitudes, to changes in practice, to student achievement improvements—could serve as a guide for when to measure what.

Fifth, working from a consistent conceptual framework could elevate professional development beyond its current craft-oriented practice to one that is based on a strong theoretical grounding and subjected to rigorous empirical scrutiny (Fishman et al., 2003). As others have argued, the goal is not to develop or validate a monolithic approach and then get others to adopt that approach. Instead, research on teacher learning ought to support adaptation and customization (Fishman & Krajcik, 2003) while maintaining a consistent core base. I offer Figure 1 as such a conceptual base, supported by theory and case-study, correlational, and experimental research.

Sixth, there has been a call for making connections between existing theories before proposing new ones (Boeler, 2002) and for developing a core theory of learning to teach (Peressini et al., 2004). The model I propose serves that purpose. Studies that are designed to test both a theory of instruction and a theory of teacher change (Wayne et al., 2008), as described earlier, have more potential to increase our understanding about how best to design professional development to improve practice (Wayne et al., 2008). Finally, working from a common research base to provide more information on when and how the elements of the model and critical features of professional development are effective will help districts and schools use the information to design and choose more effective professional development, a current pressing challenge for practitioners (Hill, 2004). My proposition follows Cochran-Smith's (2005) call for multiple empirical approaches to studying teacher education, better data collection and analysis tools, consistent use of these tools across studies, and theory-driven work where researchers build on prior studies and accumulate knowledge in a particular area.

Although quite basic, the framework I propose here provides a powerful foundation on which to build a coherent knowledge base. One example of what we can learn from studies that build on a shared foundation is a recent study by Penuel et al. (2007), which used Garet et al.'s (2001) critical features of content, time span, active learning, coherence, and work with colleagues but then extended and modified the framework to reflect a more comprehensive account of the role of context in teacher learning, by adding local supports and barriers. Penuel et al. were able to replicate the importance of the five content features and found that "the emerging research on what makes for effective professional development in science education considered broadly does provide a useful framework for examining what makes professional development effective" (p. 951); however, they found that the nature of the curricular programs made some features more or less important for supporting implementation. This study contributes an important extension of our knowledge about how quality features affect outcomes. It would not have been possible had Penuel et al. not built on previous studies and used a common conceptual framework.

The positivist model I propose is not meant to preclude, supersede, or prevent the application of complex models of teacher learning (see De Kock, Sleegers, & Voeten, 2004). We too often use the complexity and interactive nature of teacher learning as a justification for why we do not build on prior work.

Cochran-Smith and Lytle (1999) argue that there are very different views about how to define and conceptualize teacher knowledge, teacher learning, and improved practice. Although this is certainly true, I argue that we need to have a shared base from which to operate if we are to provide sensible, meaningful research-based evidence on impact that is useful for educational practitioners. For example, when determining whether a particular professional development intervention has effects on student achievement, it may be illuminating for some to know the cognitive processes through which teachers "learned" and how environmental factors interacted with the learning; but for others, knowing key features of the program that were related to outcomes is the main policy question of interest. I am arguing that both sets of questions would benefit from including the pathways and core features of professional development proposed here.

Furthermore, including the conceptual framework and core features in effectiveness studies is not meant to limit the depth at which the components might be addressed. For example, Wilson and Berne (1999) explore what "collective participation" might mean, in that they suggest the importance of understanding the features of the discourse and culture that make it productive. Pursuing such an understanding would be a useful extension of research that has established collective participation and the communities of learning that such participation fosters as features of effective professional development. From the Wilson and Berne proposition we would learn more about how and why such discourse leads to productive teacher learning. Similarly, I include context as an important component of the framework, knowing that its interpretation, emphasis, and measurement will differ depending on the study (Borko, 2004; Ensor, 2001; Grossman et al., 2000; Putnam & Borko, 2000; Shulman & Shulman, 2004).

The framework I offer does not require situating the study of teacher learning in any one perspective. Rather it could be used in studies with different perspectives on teacher learning: situated (Borko, 2004; Greeno et al., 1996; Peressini et al., 2004), cognitive (Wenger, 1987), sociocultural (Kelly, 2006), and so forth. In fact, that is one of its strengths. For example, recent work by Shulman and Shulman (2004) offers a new framework for conceptualizing teacher learning and development within communities and contexts, a framework that includes vision, motivation, understanding, practice, reflection, and community. Such a conceptualization could be used to study the effectiveness of professional development while grounding the study in the core features and links proposed here.

My propositions here are inspired by and intended to respond to and extend Borko's (2004) thoughtful analysis identifying what we know about the impact of professional development on student learning and next steps for extending knowledge, although some of the ideas might be in tension. I argue for a minimum conceptual base that we all use, whereas Borko argues for the use of multiple lenses and different conceptual frameworks. I do not believe Borko's suggestion is incongruous with my proposition, however. A careful reading of her article suggests that the heart of her suggestion is the need to focus on the individual teacher as well as the teacher's community (the sociocultural framework), and as I indicated, use of the framework I propose does not preclude differing perspectives, such as sociocultural or situated. Borko focuses on the design of the research, suggesting three phases: (a) an individual program at an individual site, (b) an individual program at multiple sites, and (c) multiple programs at multiple sites. In contrast, my thesis focuses on the use of a common conceptual framework, which could be used in studies in any of the three phases. My intention is to go beyond the scope of Borko's thesis to address the bridging of different traditions of research, for example teacher educators and policy researchers, who typically use very different analytic techniques and research designs. I am not advocating limiting either tradition but instead propose a method to integrate our growing knowledge grounded in scientifically based research, in all of its forms, and the emerging consensus on what is good professional development.

Lest the conceptual framework I propose be seen as a naïve conception of studying teacher learning that does not acknowledge the complexities, interactions, dynamics, and biases present in any research, I clearly acknowledge the positivist viewpoint that my conceptual framework represents. I believe that to make progress in understanding which features of professional development make a difference, in a way useful for educators, we can and should use empirical knowledge to drive systematic study that can build on previous work. Likely the ideas of multiple frameworks and ways of knowing, of complexity and the ambiguity of truth, are in tension with the empirical positivist view of research to find and test specific hypotheses. It is my hope that we can bridge the literature that privileges context and multiplicity over causal links, with the positivist policy evaluation literature, which privileges causal modeling over understanding how and why effects might occur. Both are critical to research and to policy. I offer the framework as a common base to allow our building on knowledge from both perspectives.

A logical challenge that follows defining professional development and establishing a core conceptual framework for studying it is figuring out how to *measure* it. I suggest that one reason why the consensus I argue for above is not more widely accepted is misconceptions about trade-offs of different ways to measure the quality of professional development and teaching. I hypothesize that the misconceptions may in part be due to folklore fostered by early validation studies that measured teachers' instruction with multiple methods, and I evaluate those early studies based on current standards of evidence and conclude that they do not provide useful results for assessing the validity of surveys, interviews, and observations for studying teacher's instruction. I then apply common dictums of methodology and measurement to studies of teacher learning, urging movement away from any automatic biases either for or against particular methods.

What Are the Implications for Modes of Inquiry in Causal Studies of Teacher Learning?

How should we measure teachers' experiences in professional development activities and subsequent changes in practice? At one time, evaluating professional development meant administering a satisfaction survey at the end of a workshop. However, we have in the past decade or more made progress on how to more usefully define and conceptualize professional development, which in turn has affected how we measure it. In addition, recently there has been considerable evolution in what the field

considers useful evidence of impact (e.g., Boruch & Mosteller, 2002; Shavelson & Towne, 2002; Slavin, 2002, 2004, 2008a). The debates regarding standards for scientific evidence and their implications for research design have been treated substantially elsewhere (Eisenhart & Towne, 2003; Maxwell, 2004; Slavin, 2008b). Here, I discuss how more rigorous evidence standards affect decisions regarding measurement and data collection in the study of teacher learning.

Whether part of an experimental, quasi-experimental, or correlational study, the challenges of measuring the quality of professional development and its effects on practice are similar. In essence, examining the effects of professional development is analogous to measuring the quality of teachers' learning experiences, the nature of teacher change, and the extent to which such change affects student learning. There is a substantial literature on measuring student learning (e.g., Cizek, Hirsch, Trent, & Crandell, 2001; Koretz, 1996; Koretz, Stecher, Klein, & McCaffrey, 1994; Linn, 1994; Mehrens, 1992), and measuring both student learning and teacher knowledge involves a host of issues related to assessment and test construction (e.g., Porter, 1998; Porter, Youngs, & Odden, 2001). Thus, here I focus on guidelines for choosing appropriate strategies for measuring professional development and its effects on changing teacher practice. Although mixed methods approaches that combine qualitative and quantitative measurement strategies hold excellent promise for transforming research in teacher learning (see Day, Sammons, & Gu, 2008; Desimone, in press; R. B. Johnson & Onwuegbuzie, 2004; Tashakkori & Teddlie, 1998), even this approach does not circumvent the inherent biases, implicit or explicit, often present in discussions surrounding empirical measurement of teacher learning and instruction.

Below, I discuss the strengths and weaknesses of what are arguably the three most commonly used and debated methods of data collection for empirical descriptive, correlational, and causal studies of teachers—observation, interviews, and surveys/questionnaires—as they pertain to measuring professional development and its effects on instruction. My purpose is not to discuss the challenges of particular research designs (e.g., Wayne et al., 2008) but rather to focus on data collection methods, which can be used across designs.

The trade-offs of alternative modes of inquiry have been discussed at length elsewhere (e.g., King, Keohane, & Verba, 1994; Ragin, 1987), but a specific discussion pertaining to professional development effects is warranted for several reasons. First, accepted tenets on the appropriateness of different modes of data collection for different purposes are not obviously applied in the literature on teacher professional development and teacher learning; instead, research is often reported with an explicit and/or implicit bias for or against particular measures. Second, specifically comparing inquiry modes in studying teaching provides context-specific guidelines. Third, a careful look at studies comparing different ways of measuring professional development and instruction provides insight into why a certain conventional wisdom about methods may have developed in the field of teacher learning. A new look at these studies with current evidentiary standards may help move us from evaluating methods by relying on conventional but often erroneous wisdom to a view more grounded in research.

Common Assumptions About Observation, Interviews, and Surveys

Common notions about the supposed strengths and weaknesses of observation, interviews, and surveys/questionnaires abound. Observation is often heralded as the most unbiased form of data collection, removing the self-report bias of surveys and interviews and allowing a clear look into what is actually occurring during a professional development activity and, subsequently, in the classroom as the teacher implements new content and strategies (Wragg, 1999). Observation is also considered the most time-consuming and expensive method of measuring professional development and teaching. Interviews are characterized as allowing for the development of a trusting relationship between the interviewer and interviewee that will elicit comprehensive and truthful information about actual implementation (Wengraf, 2004), but they are subject to interviewer bias. Surveys are lauded for being the only feasible mechanism for collecting data on large samples but are criticized for eliciting biased, socially desirable responses that overreport "good" implementation and underreport "bad" implementation.

Although each of the three methods has its supporters and detractors, survey research on professional development seems to receive the most criticism in published research. For example, implicit or explicit apologies for using survey data appear throughout the literature. A recent example relegates survey data to the status of secondhand evidence:

> Our approach to evaluation was based primarily on teacher self-report data. Given the time frame and the level of resources usually allocated to evaluations of professional development programs, there is often little opportunity to gather first-hand evidence about changes in teacher knowledge, practice, efficacy and students' learning outcomes. (Ingvarson et al., 2005, p. 18)

The argument that teachers' own assessments of their behavior is not firsthand evidence bears examination. Similarly apologetic, Fishman et al. (2003), in a review of a professional development study, say, "Yet even this study was based on teacher self-report data, not direct examination of professional development, teaching practices, or student learning" (p. 644). Borko et al. (2008) state that "self-report data have obvious limitations" (p. 418), but it is not clear that everyone in the field has similar notions of what those limitations are and how they compare with the strengths and limitations of observation and interview data. I could find no parallel examples in the professional development literature of apologies for the use of interview or observation data.

Are we justified in considering observation a more direct and unbiased method than teacher self-reports for measuring professional development and teaching? Are interviews less biased and more effective than observation and surveys in eliciting accurate information about teacher learning and its effects? In short, how different are observation, interviews, and surveys in their ability to capture the nature of professional development and the resulting changes in practice? Here, I examine early literature comparing the reliability and validity of these three approaches in the context of today's evidentiary standards, discuss more recent validity studies, and draw conclusions to extract guidelines for deciding how best

to measure professional development and teaching. I do not address the complexities of defining teaching practice in a way suitable for measurement, such as differentiating procedural and conceptual teaching (Desimone, Smith, Baker, & Ueno, 2005) and assessing the alignment of instruction to standards and assessments (Porter, Smithson, Blank, & Zeidner, 2007). Defining the teaching practices we want to measure is quite relevant to the task of including measures of teaching practice in a study of professional development effects, but a discussion of how to conceptualize teaching is beyond the scope of this article. This topic has been well addressed by others (e.g., Fenstermacher & Richardson, 2005; Kennedy, 2005, 2008; Porter, 1988).

As far as I can determine, this is the first attempt to carefully review, by today's standards, earlier validity studies of observation and self-report for measuring classroom instruction. Below, I make the case that the validity studies would not meet today's standards of scientific rigor and that their mostly negative results may have contributed to folklore in the education research community that accounts for the bias toward, or quick dismissal of, survey research to measure instruction. In this context, I emphasize the importance of adhering to well-known tenets of research design and instrumentation in evaluating the relative merits of measures of instruction. I make the argument that some of the evidence we have that supports core features is dismissed because of misconceptions about methodology.

Common Biases Against Methods for Measuring Professional Development and Its Effects on Teachers: Are They Justified?

Do observations of classroom instruction and teacher self-reports elicit the same information? A set of studies conducted in the 1960s and 1970s would lead us to say no. These early studies may have shaped the field's informal notions about how best to measure teacher learning and instruction. However, a careful look at this work suggests that we should not rely on these studies to shape our knowledge about the usefulness and comparability of different data collection instruments for measuring professional development and its effects on teaching.

Specifically, many of the early studies that may have shaped views on the comparability of classroom observation, interviews, and surveys had what would by current standards be considered fatal flaws. For example, Hook and Rosenshine (1979) reviewed a set of studies showing low correlations between classroom observations and teacher self-reports. These studies, however, did not use teacher reports of specific practices and compare them with observations of those same practices. Instead, one- or two-time observations were compared with teacher or student reports of teachers' "average" behavior (Ehman, 1970; Goodlad, Klein, & Associates, 1974). We would not expect estimations of average behavior to correlate with one or two specific observations; and in fact, the authors found little correlation.

In other studies showing differences between observations and teacher self-reports on surveys or interviews, the length or frequency of observations is unclear (Squire & Applebee, 1968; Steele, House, & Kerins, 1971; Walberg & Thomas, 1972), or there were only one or two observations for periods as short as 20 minutes (Beam & Horvath, 1975; Hardebeck, Ashbaugh, & McIntyre, 1974; Squire & Applebee, 1968). Subsequent research indicates that to elicit reliable and valid measures of teachers' overall instruction, three observations are required for one stable observation, and at least three stable observations over an extended period of time are required (Taylor, Pearson, Clark, & Walpole, 1999).

Several early studies do compare teacher and observer reports about the same lesson. However, undefined or mismatched observation protocols make these studies uninterpretable. For example, D. Johnson (1969) used different questionnaires for the observer and the teacher, thus weakening the ability to compare. And in a number of studies, teachers reported on practices before the observers observed, so the teachers were predicting what they would teach instead of reporting on how they taught (Beam & Horvath, 1975; Chall & Feldmann, 1966; Squire & Applebee, 1968; Steele et al., 1971).

More methodologically rigorous studies, in contrast, show substantive correlations between observation and self-report. The early literature shows that when self-report questions focus on a teacher's practices in a single class assignment and cover a clearly delineated and understood time frame, there is a high degree of consistency between the teacher's self-reports and an outsider's observations (Koziol & Moss, 1983; Newfield, 1980). And in retrospective reporting situations where teachers described their classroom activities after observers had carried out their observations (Hardebeck et al., 1974; Newfield, 1980), teacher–observer agreement was consistently higher.

Recent research supports these findings. Studies that use multiple observations and the exact same observer and teacher self-report protocol and that focus on behavioral rather than evaluative constructs (e.g., questions about what teachers did rather than how well they did it) show that findings from observations have moderate to high correlations with findings from surveys (Mayer, 1999; Porter, Kirst, Osthoff, Smithson, & Schneider, 1993; Ross, McDougall, Hogaboam-Gray, & LeSage, 2003). Thus, a careful look at the research shows that when teachers are reporting on concrete professional development and teaching behaviors and activities, observations and surveys can elicit much the same information.

Likewise, in comparing interviews with written or telephone surveys, research overwhelmingly suggests that both are valid forms of measurement (Cannell, Groves, Magilavy, Mathiowetz, & Miller, 1987; Groves & Kahn, 1979; Hochstim, 1967; Mangione, Hingson, & Barrett, 1982).[3] For most survey questions, information obtained by personal interview, telephone interview, and self-administered procedures have been very similar (Aquilino, 1994; Dillman & Tarnai, 1991; Tourangeau & Smith, 1998; Turner, Ku, & Rogers, 1998), even for very sensitive questions (e.g., drug use; Fowler, Roman, & Di, 1998; McHorney, Kosinski, & Ware, 1994); and for many questions this information is consistent with documented records (see Cannell, Marquis, & Laurent, 1977; Edwards, Winn, & Collins, 1996).

Thus, for behavior-based constructs, when the data collection is confidential and not linked to the teacher's own evaluation (Mayer, 1999), such as professional development activities and behavioral aspects of classroom instruction, well-constructed and administered observation, interviews, and surveys can elicit much the same information. Social desirability bias can occur in any

form of data collection. In interviews, respondents are likely to feel pressure to answer in a socially desirable way when they are face-to-face with their questioners or observers (Aquilino, 1994, 1998; Aquilino & LoSciuto, 1990; Dillman & Tarnai, 1991; Fowler, Roman, & Di, 1998; Hochstim, 1967); observers run the risk of including a rater's own biases (McCutcheon, 1981); survey respondents can have a natural positive or negative bias in how they scale their answers (Fowler, 1995); and different respondent groups, such as students, teachers, and principals, sometimes differ in their survey responses describing the same actions (Desimone, 2006; Desimone, Smith, & Frisvold, in press).

Bias depends on the quality of the instrument[4] and the type of judgment being made (Nisbett & Ross, 1980). There are extensive works on how to decrease, test, and account for observer bias (e.g., McCutcheon, 1981), interviewer bias (e.g., Rubin & Rubin, 2004), and survey bias (e.g., Sudman & Bradburn, 1982). A well-constructed and administered interview, observation, or survey protocol, when used appropriately, can provide similarly useful data, just as a poorly constructed or administered interview, observation, or survey protocol can provide skewed and biased information.

Although difficult to document, it is quite possible that conclusions drawn from early studies have shaped the field's views on the usefulness and adequacy of certain methods of data collection for studying teacher learning experiences and classroom instruction. Here I suggest that we be vigilant in not relying on conventional wisdom to shape our biases for or against certain modes of inquiry in studying teacher learning, but instead use the wealth of empirical literature we have to assess the quality of a particular mode of inquiry in a particular study, and its appropriate use.

Matching Data Collection Methods to Research Questions

Almost any methodology book or article will emphasize that research questions should drive methods. Thus, although no method should be dismissed out of hand as being inherently biased, certain methods are more appropriate than others in collecting specific information related to the effects of professional development. For example, observation and interviews are the most appropriate methods for capturing in-depth and nuanced constructs such as critical reflection and depth of focus (McLaughlin & Talbert, 2001; Putnam & Borko, 1997; Wilson & Berne, 1999), the quality of discourse and the coherence of instructional presentations (Ball & Cohen, 1999; Burstein et al., 1995), and teacher warmth, humor, and openness (Burstein et al., 1995; Rosenshine, 1979; Solomon & Kendall, 1976). Observation can make fine distinctions in teaching practice that surveys cannot make, such as distinguishing between teachers who perform reform practices perfunctorily and those who use them effectively (Cohen, 1990; Mayer, 1999; Spillane & Zeuli, 1999). Interviews and observation are also appropriate for providing narratives, examples, and anecdotes to answer research questions directed at questioning models of teacher interactions; generating hypotheses; and describing and understanding the complexities of professional development in a specific context, how beliefs and attitudes change, and the processes through which teachers change their instruction (see Merriam, 1988; Wengraf, 2004).

Survey data are by nature broad—that is both their strength and weakness. In their breadth they lack detail and complexity but gain the ability to produce statistics—quantitative, systemic numerical descriptions of events, behavior, or practice. As a result, answers to survey questions are best used to answer defined, discrete questions about frequencies and trends, specific features of professional development, and instructional time spent on specific content and practices. Teacher surveys that ask behavioral and descriptive, not evaluative, questions about the teachers' professional development experiences and teaching have been shown to have good validity and reliability (Mayer, 1999; Porter et al., 1993; Yoon, Jacobson, Garet, Birman, & Ludwig, 2004). Thus, the critical features of professional development (e.g., content focus, active learning) can be well measured with surveys. In terms of instruction, teacher surveys can provide valid and reliable data on the amount of time that teachers spend on specific practices occurring during a set time frame—up to about a year (Koziol & Moss, 1983; Mayer, 1999; Newfield, 1980). Surveys can also obtain valid and reliable data about the topic and cognitive demand coverage of a particular lesson or set of lessons (Porter, 2002).

Although surveys can do a good job of distinguishing between teachers who do and do not use reform-oriented practices (Mayer, 1999), using surveys or interviews for such a purpose is complicated and must consider teacher knowledge and beliefs (Cohen, 1990). Research shows that teachers overreport their implementation of professional development and other reforms (Cohen, 1990; Frykholm, 1996; Ross et al., 2003). Programs that aim to change teachers' behavior might instead change beliefs—and consequently self-reports—about behavior (Wubbels, Brekelmans, & Hooymaters, 1992), leaving the actual behavior unaffected. Thus, relying on self-reports of behavior might provide a too-optimistic view of the effects of a program (Wubbels et al., 1992); the same is true of relying on interviews. However, such interactions could be tested if surveys and interviews included questions designed to elicit teacher beliefs about their professional development and teaching, and the responses were then analyzed in the context of those beliefs (Cohen & Hill, 2001). Observation provides a guard against overreporting if a sufficient number of observations are implemented and the rater is well trained (Hintze & Matthews, 2004).

Moving Toward More Empirical Studies of Teacher Learning

Properly conducted observation can provide comprehensive, objective measures of what occurs in professional development and resulting classroom instruction. But observation is burdensome and expensive. Interviews, properly conducted, can provide powerfully rich explanations, examples, and hypotheses for models about how the system works. But they also exact a heavy burden on both researcher and subject and require sophisticated analytic techniques. Surveys can provide cost-effective data on discrete behavioral variables, but the data are by nature crude and limited in providing complex descriptions of professional development and teaching, or explanations for how teachers change their knowledge and skills to transform their instruction.

As research on professional development moves from its case-study base to increasingly more quantitative studies that use

surveys and more structured observation and interview protocols, we need to employ the general lessons of when and how to apply certain data collection techniques to the study of professional development. Several parallel efforts have made progress in systematizing the use of surveys to measure the quality of professional development (Yoon et al., 2004) and instruction (Porter, 2002, 2006), including the use of daily, weekly, or monthly surveys, or logs, as they are often called (Rowan, Camburn, & Correnti, 2004; Rowan, Harrison, & Hayes, 2004); using observation protocols and instruments to measure professional development delivery and implementation (Banilower & Shimkus, 2004; Garet, Yoon, & Porter, 2005; Horizon Research, Inc., 2000); and developing design strategies and instruments to link with student achievement (Yoon et al., 2004). Combining this measurement work with progress in defining and conceptualizing professional development and its effects on teaching and learning would allow us to make great strides toward improving our understanding of professional development's effects on teachers and students.

Unresolved Issues and Future Work

> I am sorry that I have had to leave so many problems unsolved. I always have to make this apology, but the world really is rather puzzling and I cannot help it.
>
> —Bertrand Russell,
> *The Philosophy of Logical Atomism,* Lecture V

In answering the overarching question "How can we best measure professional development and its effects on teachers and students?" the field of education has made considerable progress in defining what counts as professional development, as well as in delineating conceptualizations of how professional development works and how to measure various aspects of the teacher learning continuum. However, more work is needed in several key areas pertaining to these issues.

One essential step is reaching a consensus on which aspects of teacher knowledge are critical and how to measure them. Almost 10 years ago, Wilson and Berne (1999) said that "the 'what' of teacher learning needs to be identified, conceptualized, and assessed" (p. 203). Since then, we have made much progress on this front, but more needs to be done. Teacher learning may be the most difficult aspect to measure in professional development (Loucks-Horsley & Matsumoto, 1999; Wilson & Berne, 1999). Most work now acknowledges the importance of studying teacher knowledge and pedagogical content knowledge. We need more work in identifying, conceptualizing, and assessing teacher learning, including delineating the categories of knowledge that teachers should possess in a particular subject, building a conception of teacher knowledge that includes student thinking, and increasing our understanding of how teacher knowledge enables practice.

There are several schemes for categorizing and describing various types of teacher knowledge (Carter & Doyle, 1987; Clandinin, 1986; Grossman, 1990; Leinhardt & Smith, 1985) but no single agreed-upon system for characterizing either the organization of teachers' knowledge (Borko & Putnam, 1996) or appropriate aspects of such knowledge (Calderhead, 1996; Munby, Russell, & Martin, 2001). Deborah Ball and her colleagues (Ball, 2000; Hill et al., 2008; Hill, Schilling, & Ball, 2004) are currently making great progress in addressing these issues in the context of mathematics; Phelps and Schilling (2004) are at the forefront in defining and measuring content knowledge for teaching in reading; and other areas of teacher knowledge are being explored, such as technology integration in teaching (Mishra & Koehler, 2006) and science (Magnusson, Krajcik, & Borko, 1999). This work will surely be an integral part of the next decade of research on professional development.

Furthermore, measurement tools that deserve more attention are vignettes and video observations. Hamilton et al. (2003) developed and tested a vignette-based measure of reform-oriented instructional practice in mathematics, in which teachers read descriptions of various reform-oriented teaching practices and then rated the degree to which the options corresponded to their own likely behavior. The authors found that the vignette-based measure had moderate correlations with reform-oriented instruction measured by classroom observation, surveys, and logs, but concluded that the vignettes captured aspects of reform-oriented instruction not captured by other measurement methods. The authors emphasized that their vignette-based measure still had much room for improvement, especially in terms of identifying the appropriate length, level, and amount of detail. More studies comparing vignettes with other measures are needed.

Using video observation to assess both classroom instruction and teacher learning experiences has the potential to offer rich data that capture the complexity of interactions (Stigler, Gallimore, & Hiebert, 2000), but there are many challenges to address. The benefits of video as a measurement tool include obtaining a record of small-group interactions usually not possible with an observer (Borko et al., 2008) and creating a record available for multiple reexaminations with collaborative coders (Frederiksen, Sipusic, Sherin, & Wolfe, 1998; Stigler et al., 2000; Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999; Ulewicz & Beatty, 2001). Challenges include questions about how to analyze videos for research purposes, the role that protocol and a conceptual framework play in how videos get analyzed (Sherin & Han, 2004), addressing privacy and confidentiality issues, capturing whatever in the context of a lesson is important for interpreting what is seen in the video (Erickson, 1986), and determining what aspects of teaching should be videotaped (Hall, 2000; Ulewicz & Beatty, 2001). Recent work has focused on how using video for classroom and teacher learning can expand our knowledge and create descriptive, explanatory, or expository accounts of learning and teaching (e.g., Goldman & McDermott, 2007; Miller, 2007); however, there remain challenges in connection with employing this medium to its maximum potential and determining how to use it in relation to other measurement tools. Exploring the logistics and practicality of using video observation to assist in our understanding of professional development and its effects is an obvious next step.

Another issue is that, although there is a consensus on at least some of the features of high-quality professional development, we do not yet have a clear indication of thresholds for these features. For example, how much professional development is enough? More than a decade ago, Stout (1996) said that "no evidence exists to allow a sensible policy decision about the amount of staff development

needed to accomplish any given purpose" (p. 6); we still have little guidance on this issue. An exploration of the question would require comparing activities with the same content but varying amounts of time and then relating amounts of time in the professional development to the extent of changes in instruction and student learning. Similarly, other key characteristics, such as active learning, might be studied to provide insights about threshold levels required to elicit desired change.

Still another area for future study is professional development using nonvolunteers. Where professional development research with a focus on teacher and student learning has been conducted (e.g., Carpenter, Fennema, & Franke, 1996; Marx, Freeman, Krajcik, & Blumenfeld, 1997), it has focused on groups of volunteer teachers who are motivated to change or to try something new (Supovitz & Zeif, 2000), and there is evidence that the most qualified teachers are the ones who seek out professional development with effective features such as content focus (Desimone, Smith, & Ueno, 2006). Limited evidence suggests that policy can play a role in influencing who participates in effective professional development (Desimone, Smith, & Rowley, 2007) and that teachers increasingly are taking part in more effective professional development (T. M. Smith & Desimone, 2003); however, it is still unclear how findings might be different if effectiveness studies were not limited to motivated volunteers (Bobrowsky, Marx, & Fishman, 2001).

In addition, in the vein of working for consistent definitions of professional development and a core conceptual framework, we would benefit from developing a consistent set of instruments to measure professional development. Currently, there are a few promising examples of sets of data collection tools designed for this purpose (Banilower & Shimkus, 2004; Garet et al., 2005; Horizon Research, Inc., 2000; P. S. Smith, 2005), but they have not been subject to repeated use and validation and are not widely available (Borko, 2004). Such tools would go a long way in contributing to the increasing numbers of experimental designs and multiple-site longitudinal studies currently under way. A first step toward developing a consistent, valid set of instruments to potentially use across studies (P. S. Smith, 2005) would be to require the reporting of measures used in professional development studies, a practice that is now rare (e.g., Ingvarson et al., 2005).

Finally, we need more work that links professional development and changes in teaching practice to student achievement. Although it has been repeatedly noted that it is difficult and expensive to design and conduct evaluations that isolate and measure the specific effects of professional development on student achievement (Desimone, Porter, et al., 2002; Fishman et al., 2003; Frechtling et al., 1995; Killion, 1998), several such studies are currently under way or have recently been completed (e.g., Garet et al., 2008; Glazerman et al., 2008), indicating that these studies are feasible.

Conclusion: Toward Better Conceptualizations and Measures of Professional Development for Research and Policy

> It is never too late to become reasonable and wise; but if the knowledge comes late, there is always more difficulty in starting a reform.
>
> —Immanuel Kant,
> *Prolegomena to Any Future Metaphysics*

In answering the broad question of how best to measure professional development and its effects on instruction and student achievement, I make several propositions. I argue that as a field we have reached an empirical consensus on a set of core features and a conceptual framework for teacher learning, and that we should use the framework in future studies of the effectiveness of professional development while allowing for individual adaptation. These points of consensus would serve as a guide for what is essential to measure, and allow comparison across studies, to build our knowledge base. I link reticence about acknowledging a consensus at least in part to conventional wisdom about methods for studying professional development and teaching. I point to flaws in early validation studies that may have contributed to unfounded bias against certain types of measurement. Finally, I recommend that we move away from any automatic biases either for or against observation, interviews, or surveys, and instead base our evaluations and critiques of measurement instruments on the quality of their design and administration, according to best practice, and on their appropriateness given a study's particular research question.

Professional development is a key to reforms in teaching and learning, making it essential that we use best practice to measure its effects. Several decades of research have provided us with a wealth of information to improve our conceptualizations and measures of professional development. I suggest that we take better advantage of this research to elevate the quality of professional development studies and subsequently to elevate our understanding of how best to shape and implement teacher learning opportunities for the maximum benefit of both teachers and students.

NOTES

[1]For comprehensive discussions of the research literature on teacher learning and professional development, see Borko and Putnam (1996), Putnam and Borko (1997), and Wilson and Berne (1999).

[2]Of course, much can be done to prepare teachers in their preservice training. That is beyond the scope of this article, although it has been the focus of much other work (see Wilson, Floden, & Ferrini-Mundy, 2002).

[3]For good summaries of results of studies comparing different forms of measurement, see de Leeuw and van der Zouwen (1988) and Dillman (2000).

[4]For detailed discussion of how to produce a good survey instrument, see Converse and Presser (1986), Fowler (1995), and Sudman and Bradburn (1982); for interviews, see Wengraf (2004); and for observation, see Wragg (1999). The lessons as related to the study of teacher learning are that achieving interrater reliability is essential for sound observation (Hintze & Matthews, 2004; S. Lee, 2005; Walsh, 1967), and the wording, format, and choice of response categories strongly influence self-reports of behavior and attitudes in interviews and surveys (Fowler, 2002; Schwartz, 1999; Sudman, Bradburn, & Schwarz, 1996); respondents draw on formal features of a research instrument, such as the nature of preceding questions (Moxey & Sanford, 1992), to disambiguate the meaning of the questions posed to them (Schwartz, 1999). Analogous to establishing interrater reliability, interviews and surveys should undergo proper development and piloting, which may include focus groups, critical review, cognitive laboratory interviews, and pretesting (Fowler, 2002), with particular attention to teacher interpretations of language (Desimone & LeFloch, 2004; Ross, McDougall, Hogaboam-Gray, & LeSage, 2003; Tourangeau, 1984). In addition, many studies

have shown that composite survey measures are more reliable than single-item indicators (Bennet, 1976; Hook & Rosenshine, 1979; Light, Singer, & Willett, 1990; Mayer, 1999; Stallings & Kaskowitz, 1974).

REFERENCES

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5–11.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1997). Situated versus cognitive perspectives: Form versus substance. *Educational Researcher, 26*(1), 18–21.

Angrist, J., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics, 19*(2), 343–369.

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. *Public Opinion Quarterly, 58*(2), 210–240.

Aquilino, W. S. (1998). Effects of interview mode on measuring depression in younger adults. *Journal of Official Statistics, 14*(1), 15–29.

Aquilino, W. S., & LoSciuto, L. (1990). Effects of interview mode on self-reported drug use. *Public Opinion Quarterly, 54,* 362–395.

Ball, D. L. (1996). Teacher learning and the mathematics reforms: What we think we know and what we need to learn. *Phi Delta Kappan, 77,* 500–508.

Ball, D. L. (2000). Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. *Journal of Teacher Education, 51,* 241–247.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher, 25*(9), 6–8.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3–32). San Francisco: Jossey-Bass.

Banilower, E., Heck, D., & Weiss, I. (2005). Can professional development make the vision of the standards a reality? The impact of the National Science Foundations Local Systemic Change Through Teacher Enhancement Initiative. *Journal of Research in Science Teaching, 44*(3), 375–395.

Banilower, E., & Shimkus, E. (2004). *Professional development observation study.* Chapel Hill, NC: Horizon Research.

Beam, K., & Horvath, R. (1975). Differences among teachers' and students' perceptions of science classroom behaviors, and actual classroom behaviors. *Science Education, 59,* 333–344.

Bennet, N. (1976). *Teaching styles and pupil program.* London: Open Books.

Birman, B., Le Floch, K. C., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., et al. (2007). *State and local implementation of the No Child Left Behind Act: Vol. 2. Teacher quality under NCLB: Interim report.* Washington, DC: U.S. Department of Education; Office of Planning, Evaluation and Policy Development; Policy and Program Studies Service.

Blasé, J., & Blasé, J. (2000). Effective instructional leadership: Teachers' perspectives on how principals promote teaching and learning in schools. *Journal of Educational Administration, 38,* 130–141.

Bobrowsky, W., Marx, R., & Fishman, B. (2001, March). *The empirical base for professional development in science education: Moving beyond volunteers.* Paper presented at the annual meeting of the National Association of Research on Science Teaching, St. Louis, MO.

Boeler, J. (2002). Exploring the nature of mathematical activity: Using theory, research and "working hypotheses" to broaden conceptions of mathematics knowing. *Educational Studies in Mathematics, 51*(1–2), 3–21.

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

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discourse in mathematics professional development. *Teaching and Teacher Education, 24,* 417–436.

Borko, H., & Putnam, R. (1995). Expanding a teachers' knowledge base: A cognitive psychological perspective on professional development. In T. Guskey & M. Huberman (Eds.), *Professional development in education: New paradigms and practices* (pp. 35–66). New York: Teachers College Press.

Borko, H., & Putnam, R. (1996). Learning to teach. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 673–708). New York: Macmillan.

Boruch, R. F., & Mosteller, F. (Eds.). (2002). *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institution Press.

Bressoux, P. (1996). The effect of teachers' training of pupils' achievement: The case of elementary schools in France. *School Effectiveness and School Improvement, 7*(3), 252–279.

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., & Guitton, G. (1995). *Validating national curriculum indicators.* Santa Monica, CA: RAND.

Calderhead, J. (1996).Teachers: Beliefs and knowledge. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 709–725). New York: Macmillan.

Cannell, C., Groves, R., Magilavy, L., Mathiowetz, N., & Miller, P. (1987). *An experimental comparison of telephone and personal health surveys.* Washington, DC: National Center for Health Statistics.

Cannell, C., Marquis, K., & Laurent, A. (1977). A summary of studies of interviewing methodology. *Vital and Health Statistics, 2*(69), 1–78.

Carey, N., & Frechtling, J. (1997, March). *Best practice in action: Follow-up survey on teacher enhancement programs.* Arlington, VA: National Science Foundation.

Carnegie Forum on Education and the Economy. (1986). *A nation prepared: Teachers for the 21st century.* New York: Carnegie Corporation.

Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *Elementary School Journal, 97,* 3–20.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*(4), 499–531.

Carpenter, T. P., Franke, M. L., & Levi, L. (1998, April). *Teachers' epistemological beliefs about their knowledge of children's mathematical thinking.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Carter, K., & Doyle, W. (1987). Teachers' knowledge structures and comprehension processes. In J. Calderhead (Ed.), *Exploring teachers' thinking* (pp. 147–160). London: Cassell.

Chall, J., & Feldmann, S. (1966). First-grade reading: An analysis of the interaction of professed methods, teacher implementation, and child background. *Reading Teacher, 19,* 569–575.

Cizek, G., Hirsch, T., Trent, E., & Crandell, J. (2001). A preliminary investigation of pupil proficiency testing and state education reform initiatives. *Educational Assessment, 7*(4), 283–302.

Clandinin, D. J. (1986). *Classroom practice: Teacher images in practice.* London: Falmer.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18,* 947–967.

Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. *Educational Researcher, 23*(7), 13–19.

Cochran-Smith, M. (2005). Studying teacher education: What we know and need to know. *Journal of Teacher education, 56*(4), 301–306.

Cochran-Smith, M., & Lytle, S. I. (1999). Relationships of knowledge and practice: Teacher learning in communities. *Review of Research in Education, 24,* 249–305.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*(3), 311–329.

Cohen, D. K., & Ball, D. (1990). Policy and practice: An overview. *Educational Evaluation and Policy Analysis, 12*(3), 347–353.

Cohen, D. K., & Hill, H. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*(2), 294–343.

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

Consortium for Policy Research in Education. (1998). *A close look at the effects on classroom practice and student performance: A report of the fifth year of the Merck Institute for Science Education* (CPRE Evaluation Report). Philadelphia: Author.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* New Delhi, India: Sage.

Darling-Hammond, L. (1993). Reframing the school reform agenda: Developing capacity for school transformation. *Phi Delta Kappan, 74,* 735–761.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching.* New York: National Commission on Teaching and America's Future.

Darling-Hammond, L., & McLaughlin, M. (1999). Investing in teaching as a learning profession: Policy problems and prospects. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 376–412). San Francisco: Jossey-Bass.

Darling-Hammond, L., & Sykes, G. (Eds.). (1999). *Teaching as the learning profession: Handbook of policy and practice.* San Francisco: Jossey-Bass.

Day, C., Sammons, P., & Gu, Q. (2008). Combining qualitative and quantitative methodologies in research on teachers' lives, work, and effectiveness: From integration to synergy. *Educational Researcher, 37*(6), 330-342.

De Kock, A., Sleegers, P., & Voeten, M. J. (2004). New learning and the classification of learning environments in secondary education. *Review of Educational Research, 74*(2), 141–170.

De Leeuw, E. D., & van der Zouwen, J. (1988). Data quality in telephone and face-to-face surveys: A comparative meta-analysis. In R. M. Groves, P. P. Biemer, I. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283–300). New York: John Wiley.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2002). *The qualitative inquiry reader.* Thousand Oaks, CA: Sage.

Desimone, L. M. (2002). What makes Comprehensive School Reform successful? *Review of Educational Research, 72*(3), 433–479.

Desimone, L. M. (2003). Toward a more refined theory of school effects: A study of the relationship between professional community and mathematic teaching in early elementary school. In C. Miskel & W. Hoy (Eds.), *Research and theory in educational administration.* Greenwich, CT: Information Age.

Desimone, L. M. (2006). Consider the source: Response differences among teachers, principals and districts on survey questions about their education policy environment. *Educational Policy, 20*(4), 640–676.

Desimone, L. M. (in press). Complementary methods for policy research. In D. Plank, G. Sykes, & B. Schneider (Eds.), *AERA handbook on education policy research.* Washington, DC: American Educational Research Association.

Desimone, L. M., Birman, B., Porter, A., Garet, M., & Yoon, K. S. (2003). Improving teachers' in-service professional development in mathematics and science: The role of postsecondary institutions. *Educational Policy, 17*(5), 613–649.

Desimone, L. M., Garet, M., Birman, B., Porter, A., & Yoon, K. S. (2002). How do district management and implementation strategies relate to the quality of the professional development that districts provide to teachers? *Teachers College Record, 104*(7), 1265–1312.

Desimone, L. M., & LeFloch, K. (2004). Probing the "trickle down" effect of standards and assessments: Are we asking the right questions? *Educational Evaluation and Policy Analysis, 26*(1), 1–22.

Desimone, L. M., Porter, A. C., Garet, M., Yoon, K. S., & Birman, B. (2002). Does professional development change teachers' instruction? Results from a three-year study. *Educational Evaluation and Policy Analysis, 24*(2), 81–112.

Desimone, L. M., Smith, T., Baker, D., & Ueno, K. (2005). The distribution of teaching quality in mathematics: Assessing barriers to the reform of United States mathematics instruction from an international perspective. *American Educational Research Journal, 42*(3), 501–535.

Desimone, L. M., Smith, T., & Frisvold, D. (2007). Is NCLB increasing teacher quality for students in poverty? In A. Gamoran (Ed.), *Standards-based and the poverty gap: Lessons from No Child Left Behind* (pp. 89–119). Washington, DC: Brookings Institution Press.

Desimone, L. M., Smith, T., & Frisvold, D. (in press). How similar are student and teacher reports of classroom instruction? *Educational Policy.*

Desimone, L. M., Smith, T. M., Hayes, S., & Frisvold, D. (2005). Beyond accountability and average math scores: Relating multiple state education policy attributes to changes in student achievement in procedural knowledge, conceptual understanding and problem solving in mathematics. *Educational Measurement: Issues and Practice, 24*(4), 5–18.

Desimone, L. M., Smith, T. M., & Rowley, K. (2007). Does policy influence mathematics and science teachers'participation in professional development? *Teachers College Record, 109*(5), 1086–1122.

Desimone, L. M., Smith, T., & Ueno, K. (2006). Are teachers who need sustained, content-focused professional development getting it? An administrator's dilemma. *Educational Administration Quarterly, 42*(2), 179–215.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York: John Wiley.

Dillman, D. A., & Tarnai, J. (1991). Mode effects of cognitively designed recall questions: A comparison of answers to telephone and mail surveys. In P. N. Beimer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 367–393). New York: John Wiley.

Edwards, W. S., Winn, D. M., & Collins, J. G. (1996). Evaluation of 2-week doctor visit reporting in the National Health Interview Survey. *Vital Health Statistics*, 2nd ser. (122), 1–46. Hyattsville, MD: National Center for Health Statistics.

Ehman, L. (1970). *A comparison of three sources of classroom data: Teachers, students, and systematic observation.* Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, MN.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher, 32,* 31–38.

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

Elmore, R. F., & Burney, D. (1996, March). *Staff development and instructional improvement: Community District 2, New York City.* Philadelphia: Consortium for Policy Research in Education.

Elmore, R. F., & Burney, D. (1997). *Investing in teacher learning: Staff development and instructional improvement in Community School District #2, New York City.* New York: National Commission on Teaching and America's Future. (ERIC Document Reproduction Service No. ED 416203)

Ensor, P. (2001). From preservice mathematics teacher education to beginning teaching: A study in recontextualizing. *Journal for Research in Mathematics Education, 32,* 296–320.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119–161). New York: Macmillan.

Feiman-Nemser, S. (1985). Learning to teach. In L. Shulman & G. Sykes (Eds.), *Handbook of teaching and policy* (pp. 150–170). New York: Longman.

Fenstermacher, G., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record, 107*(1), 186–213.

Firestone, W. (1996). Images of teaching and proposals for reform: A comparison of ideas from cognitive and organizational research. *Educational Administration Quarterly, 32*(2), 209–232.

Firestone, W., Mangin, M., Martinez, M., & Polovsky, T. (2005). Leading coherent professional development: A comparison of three districts. *Educational Administration Quarterly, 41*(3), 413–448.

Fishman, B., & Krajcik, J. S. (2003). What does it mean to create sustainable science curriculum innovations? *Science Education, 87*(4), 564–573.

Fishman, J. J., Marx, R. W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and Teacher Education, 19,* 643–658.

Fowler, F. J., Jr. (1995). Improving survey questions: Design and evaluation. In *Applied social research methods series* (Vol. 38). Thousand Oaks, CA: Sage.

Fowler, F. J., Jr. (2002). Survey research methods. In *Applied social research methods series* (3rd ed., Vol. 1). Thousand Oaks, CA: Sage.

Fowler, F. J., Jr., Roman, A., & Di, Z. (1998). Mode effects in a survey of Medicare prostate surgery patients. *Public Opinion Quarterly, 62,* 29–46.

Franke, M. L., Carpenter, T. P., & Levi, L. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal, 38,* 653–689.

Frechtling, J., Sharp, L., Carey, N., & Vaden-Kiernan, N. (1995). *Teacher enhancement programs: A perspective on the last four decades* [Monograph]. Arlington, VA: National Science Foundation Directorate for Education and Human Resources.

Frederiksen, J., Sipusic, M., Sherin, M., & Wolfe, E. (1998). Video portfolio assessment: Creating a framework for viewing the functions of teaching. *Educational Assessment, 5*(4), 225–297.

Frykholm, J. A. (1996). Pre-service teachers in mathematics: Struggling with the standards. *Teaching and Teacher Education, 12*(6), 665–681.

Fullan, M. (1991). *The new meaning of educational change.* New York: Teachers College Press.

Fullan, M. (1993). *Change forces: Probing the depth of educational reform.* New York: Falmer.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008–4030). Report prepared for the Institute of Education Sciences. Washington, DC: Institute of Education Sciences.

Garet, M. S., Porter, A. C., Desimone, L. M., Birman, B., & Yoon, K. S. (2001). What makes professional development effective? Analysis of a national sample of teachers. *American Educational Research Journal, 38*(3), 915–945.

Garet, M. S., Yoon, K. S., & Porter, A. P. (2005, April). *Measuring differences in the quality of professional development.* Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.

Gearhart, M., & Wolf, S. A. (1994). Engaging teachers in assessment of their students' writing: The role of subject matter knowledge. *Assessing Writing, 1,* 67–90.

Glazerman, S., Dolfin, S., Bleeker, M., Johnson, A., Isenberg, E., Lugo-Gil, J., et al. (2008). *Impacts of comprehensive teacher induction: Results from the first year of a randomized control study* (NCEE 2009–4034). Washington, DC: U.S. Department of Education.

Goldman, S., & McDermott, R. (2007). Staying the course with video analysis. In R. Goldman, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 101–114). Mahwah, NJ: Lawrence Erlbaum.

Goodlad, J., Klein, M., & Associates. (1974). *Looking behind the classroom door: A useful guide to observing schools in action.* Worthington, OH: Charles A. Jones.

Greenleaf, C. L., Schoenbach, R., Cziko, C., & Mueller, F. L. (2001). Apprenticing adolescent readers to academic literacy. *Harvard Educational Review, 71*(1), 79–129.

Greeno, J. G. (1997). On claims that answer the wrong questions. *Educational Researcher, 26*(1), 5–17.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15–46). New York: Simon & Schuster Macmillan.

Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education.* New York: Teachers College Press.

Grossman, P. L., Valencia, S. W., Evans, K., Thompson, C., Martin, S., & Plaoe, N. (2000). Transitions into teaching: Learning to teach writing in teacher education and beyond. *Journal of Literacy Research, 32*(4), 631–662.

Grossman, P. L., Wineburg, S., & Woolworth, S. (2001). Toward a theory of teacher community. *Teachers College Record, 103,* 942–1012.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews.* New York: Academic Press.

Guskey, T. R. (1994). Results-oriented professional development: In search of an optimal mix of effective practices. *Journal of Staff Development, 15*(4), 42–50.

Guskey, T. R. (2000). *Evaluating professional development.* Thousand Oaks, CA: Corwin Press.

Guskey, T. R. (2002). Does it make a difference? Evaluating professional development. *Educational Leadership, 59*(6), 45–51.

Guskey, T. R., & Sparks, D. (1991). What to consider when evaluating staff development. *Educational Leadership, 49*(3), 73–76.

Hall, R. (2000). Videorecording as theory. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 647–664). Mahwah, NJ: Lawrence Erlbaum.

Hamilton, L., McCaffrey, D., Stecher, B., Klein, S., Robyn, A., & Bugliari, D. (2003). Studying large-scale reforms of instructional practice: An example from mathematics and science. *Educational Evaluation and Policy Analysis, 25*(1), 1–29.

Hardebeck, R., Ashbaugh, C., & McIntyre, K. (1974). *Individualization of instruction by vocational and non-vocational teachers: Self-reports*

compared with observations. Austin: University of Texas. (ERIC Document Reproduction Service No. ED 131202)

Hawley, W., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp.127–150). New York: Teachers College Press.

Heck, D. J., Banilower, E. R., Weiss, I. R., & Rosenberg, S. L. (2008). Studying the effects of professional development: The case of the NSF's local systemic change through teacher enhancement initiative. *Journal for Research in Mathematics Education, 39*(2), 113–152.

Hill, H. C. (2004). Professional development standards and practices in elementary school mathematics. *Elementary School Journal, 104*(3), 215–231.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372–400.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*(1), 11–30.

Hintze, J., & Matthews, W. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*(2), 258–270.

Hochstim, J. R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association, 62,* 976–987.

The Holmes Group. (1986). *Tomorrow's teachers: A report of the Holmes Group.* East Lansing, MI: Author.

Hook, C. M., & Rosenshine, B. V. (1979). Accuracy of teacher reports of their classroom behavior. *Review of Educational Research, 49*(1), 1–12.

Horizon Research, Inc. (2000). *Core evaluation manual: Professional development observation protocol.* Chapel Hill: Author.

Ingvarson, L., Meiers, M., & Beavis, A. (2005). Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes and efficacy. *Education Policy Analysis Archives, 13*(10). Retrieved May 23, 2005, from http://epaa.asu.edu/epaa/v13n10/

Jacob, B., & Lefgren, L. (2004). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources, 39*(1), 50–79.

Jeanpierre, B., Oberhauser, K., & Freeman, C. (2005). Characteristics of professional development that effect change in secondary science teacher's classroom practices. *Journal of Research in Science Teaching, 42*(6), 668–690.

Johnson, C., Kahle, J., & Fargo, J. (2007). A study of the effect of sustained, whole-school professional development on student achievement in science. *Journal of Research in Science Teaching, 44*(6), 775–786.

Johnson, D. (1969). *The relationship of self-supervision to change in selected attitudes and behaviors.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33,* 14–26.

Kelly, P. (2006). What is teacher learning? A socio-cultural perspective. *Oxford Review of Education, 32*(4), 505–519.

Kennedy, M. M. (1998). *Form and substance in in-service teacher education* (Research Monograph No. 13). Arlington, VA: National Science Foundation.

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

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 59–63.

Killion, J. (1998). Scaling the elusive summit. *Journal of Staff Development, 19*(4), 12–16.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research.* Princeton, NJ: Princeton University Press.

Koretz, D. (1996). Using student assessments for educational accountability. In E. Hanushek & D. Jorgenson (Eds.), *Improving American's schools: The role of incentives* (pp. 171–195). Washington, DC: National Research Council.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Koziol, S., Jr., & Moss, P. (1983). *The use of elementary students' reports to assess the accuracy of teacher self-reports.* Paper presented at the annual meeting of the American Educational Research Association, Montréal, Quebec.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Lee, O., Deaktor, R., Enders, C., & Lambert, J. (2008). Impact of a multiyear professional development intervention on science achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching, 45*(6), 726–747.

Lee, S. (Ed.). (2005). *Encyclopedia of school psychology.* Thousand Oaks, CA: Sage.

Leinhardt, G., & Smith, D. A. (1985). Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology, 77,* 247–271.

Lieberman, A. (1995). Restructuring schools: The dynamics of changing practice, structure, and culture. In A. Lieberman (Ed.), *The work of restructuring schools: Building from the ground up* (pp. 1–17). New York: Teachers College Press.

Lieberman, A. (1996). Creating intentional learning communities. *Educational Leadership, 54*(3), 51–55.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education.* Cambridge, MA: Harvard University Press.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.

Little, J. W. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal, 19*(3), 325–340.

Little, J. W. (1987). Teachers as colleagues. In V. Richardson-Koehler (Ed.), *Educators' handbook: A research perspective* (pp. 491–518). New York: Longman.

Little, J. W. (1993). Teachers' professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis, 15*(2), 129–151.

Little, J. W. (1999). Organizing schools for teacher learning. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 233–262). San Francisco: Jossey-Bass.

Little, J. W. (2002). Locating learning in teachers' communities of practice: Opening up problems of analysis in records off everyday work. *Teaching and Teacher Education, 18,* 917–946.

Loucks-Horsley, S., Harding, C., Arbuckle, M., Murray, L., Dubea, C., & Williams, M. (1987). *Continuing to learn: A guidebook for teacher development.* Andover, MA, and Oxford, OH: The Regional

Laboratory for Educational Improvement of the Northeast and Islands and the National Staff Development Council.

Loucks-Horsley, S., Hewson, P. W., Love, N., & Stiles, K. (1998). *Designing professional development for teachers of science and mathematics.* Thousand Oaks, CA: Corwin Press.

Loucks-Horsley, S., Love, N., Stiles, K. E., & Mundry, S. E., & Hewson, P. (2003). *Designing professional development for teachers of science and mathematics.* Thousand Oaks, CA: Corwin Press.

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*(5), 258–271.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge: The construct and its implications for science education* (pp. 95–132). Dordrecht, the Netherlands: Kluwer.

Mangione, T. W., Hingson, R., & Barrett, J. (1982). Collecting sensitive data: A comparison of three survey strategies. *Sociological Methods and Research, 10*(3), 337–346.

Marx, R. W., Freeman, J. G., Krajcik, J. S., & Blumenfeld, P. C. (1997). Professional development of science teachers. In B. Fraser & K. Tobin (Eds.), *International handbook of science education* (pp. 667–680). Dordrecht, the Netherlands: Kluwer.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3–11.

Mayer, D. P. (1998). Do new teaching standards undermine performance on old tests? *Educational Evaluation and Policy Analysis, 20,* 53–78.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29–45.

McCutcheon, G. (1981). On the interpretation of classroom observations. *Educational Researcher, 10*(5), 5–10.

McHorney, C. A., Kosinski, M., & Ware, J. E., Jr. (1994). Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: Results from a national survey. *Medical Care, 32,* 551–567.

McLaughlin, M. W., & Talbert, J. (1993). Introduction: New visions of teaching. In D. Cohen, M. McLaughlin, & J. Talbert (Eds.), *Teaching for understanding: Challenges for policy and practice* (pp. 1–12). San Francisco: Jossey-Bass.

McLaughlin, M. W., & Talbert, J. E. (2001). *Professional communities and the work of high school teachers.* Chicago: University of Chicago Press.

Mehrens, W. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3–9.

Merriam, S. (1988). *Case-study research in education: A qualitative approach.* San Francisco: Jossey-Bass.

Miles, M., & Huberman, A. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). London: Sage.

Miller, K. (2007). Learning from classroom video: What makes it compelling and what makes it hard. In R. Goldman, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 321–334). Mahwah, NJ: Lawrence Erlbaum.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054.

Moxey, L. M., & Sanford, A. J. (1992). Context effects and the communicative functions of quantifiers: Implications for their use in attitude research. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 279–296). New York: Springer Verlag.

Munby, H., Russell, T., & Martin, A. K. (2001). Teachers' knowledge and how it develops. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 877–904). Washington, DC: American Educational Research Association.

National Commission on Teaching and America's Future. (1997). *Doing what matters most: Investing in quality teaching.* New York: Author.

Newfield, J. (1980). Accuracy of teacher reports: Reports and observations of specific classroom behaviors. *Journal of Educational Research, 74*(2), 78–82.

Newmann, F., & Associates (1996). *Authentic achievement: Restructuring schools for intellectual quality.* San Francisco: Jossey-Bass.

Nir, A., & Bogler, R. (2008). The antecedents of teacher satisfaction with professional development programs. *Teaching and Teacher Education, 24,* 377–386.

Nisbett, R., & Ross, L. (1980) *Human inference: Strategies and social shortcomings of social judgment.* London: Prentice Hall.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, § 9101, 34.

Penuel, W. R., Fishman, B., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal, 44*(4), 921–958.

Peressini, D., Borko, H., Romagnano, L., Knuth, E., & Willis, C. (2004). A conceptual framework for learning to teach secondary mathematics: A situative perspective. *Educational Studies in Mathematics, 56*(1), 67–96.

Phelps, G., & Schilling, S. (2004). Developing measures of content knowledge for teaching reading. *Elementary School Journal, 105*(1), 31–48.

Porter, A. C. (1988). Understanding teaching: A model for assessment. *Journal of Teacher Education, 39*(4), 2–7.

Porter, A. C. (1989). External standards and good teaching: The pros and cons of telling teachers what to do. *Educational Evaluation and Policy Analysis, 11*(4), 343–356.

Porter, A. C. (1994). National standards and school improvement in the 1990s: Issues and promise. *American Journal of Education, 102,* 421–449.

Porter, A. C. (1998). Dilemmas in assessing academic achievement. In N. M. Lambert & B. L. McCombs (Eds.), *How students learn: Reforming schools through learner-centered education* (pp. 339–350). Washington, DC: American Psychological Association.

Porter, A. C. (2002, October). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3–14.

Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–159). Washington, DC: American Educational Research Association.

Porter, A. C., Kirst, M. W., Osthoff, E. J., Smithson, J. L., & Schneider, S. A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms.* New Brunswick, NJ: Consortium for Policy Research in Education.

Porter, A. C., Smithson, J. L., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*(1), 27–51.

Porter, A. C., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 259–297). Washington, DC: American Educational Research Association.

Putnam, R. T., & Borko, H. (1997). Teacher learning: Implications of new views of cognition. In B. J. Biddle, T. L. Good, & I. F. Goodson (Eds.), *International handbook of teachers and teaching* (2nd ed., pp. 1223–1296). Dordrecht, the Netherlands: Kluwer.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4–15.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies.* Berkeley: University of California Press.

Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research, 75*(2), 211–246.

Richardson, V. (1996). The role of attitudes and beliefs in learning to teach. In J. Sikula, T. Buttery, & E. Guyton (Eds.), *Handbook of research on teacher education* (pp. 102–119). New York: Simon & Schuster Macmillan.

Rosenholtz, S. J. (1989). Workplace conditions that affect teacher quality and commitment: Implications for teacher induction programs. *Elementary School Journal, 89*(4), 421–439.

Rosenholtz, S. J. (1991). *Teachers' workplace: The social organization of schools.* New York: Teachers College Press.

Rosenshine, B. (1979). Content, time and direct instruction. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching* (pp. 28–56). Berkeley, CA: McCutchan.

Ross, J. A., McDougall, D., Hogaboam-Gray, A., & LeSage, A. (2003). A survey measuring elementary teachers' implementation of standards-based mathematics teaching. *Journal for Research in Mathematics Education, 34*(1), 344–363.

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in the third-grade classrooms. *Elementary School Journal, 105*(1), 75–101.

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *Elementary School Journal, 105*(1), 103–127.

Rubin, H., & Rubin, I. (2004). *Qualitative interviewing: The art of hearing data* (2nd ed.). Thousand Oaks, CA: Sage.

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

Schifter, D., & Fosnot, C. (1993). *Reconstructing mathematics education: Stories of teachers meeting the challenge of reform.* New York: Teachers College Press.

Schultz, K., Jones-Walker, C. E., & Chikkatur, A. P. (2008). Listening to students, negotiating beliefs: Preparing teachers for urban classrooms. *Curriculum Inquiry, 38*(2), 155–187.

Schwab, T. (1973). The practical translation into curriculum. *School Review, 81,* 501–522.

Schwartz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93–105.

Scribner, J. (1999). Professional development: Untangling the influence of work context on teacher learning. *Educational Administration Quarterly, 35*(2), 238–266.

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

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

Shulman, L. S., & Shulman, J. H. (2004). How and what teachers learn: Shifting perspective. *Journal of Curriculum Studies, 36*(2), 256–271.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31,* 15–21.

Slavin, R. E. (2004). Education research can and must address "What works" questions. *Educational Researcher, 33,* 27–28.

Slavin, R. E. (2008a). Evidence-based reform in education: Which evidence counts? *Educational Researcher, 37,* 47–50.

Slavin, R. E. (2008b). Perspectives on evidence-based research in education—What works? Issues in Synthesizing Educational Program Evaluations. *Educational Researcher, 37,* 5–14.

Smith, P. S. (2005). *Assessing teacher learning about science teaching (ATLAST)* (EHR-0335328). Report of Project Activities and Findings. Year Two. Chapel Hill, NC: Horizon Research.

Smith, T. M., & Desimone, L. M. (2003). Do changes in patterns of participation in teachers' professional development reflect the goals of standards-based reform? *Educational Horizons, 81*(3), 119–129.

Smith, T. M., Desimone, L. M., & Ueno, K. (2005). "Highly qualified" to do what? The relationship between NCLB teacher quality mandates and the use of reform-oriented instruction in middle school math. *Educational Evaluation and Policy Analysis, 27*(1), 75–109.

Smith, T. M., Desimone, L. M., Zeidner, T., Dunn, A. C., Bhatt, M., & Rumyantseva, N. (2007). Inquiry-oriented instruction in science: Who teachers that way? *Educational Evaluation and Policy Analysis, 9*(29), 169–199.

Snow, C. E., Burns, S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* National Research Council. Washington, DC: National Academy Press.

Solomon, D., & Kendall, A. (1976). *Final report: Individual characteristics and children's performance in varied educational settings.* Rockville, MD: Montgomery County Public Schools.

Spillane, J. P. (2004). *Standards deviations: How local schools misunderstand policy.* Cambridge, MA: Harvard University Press.

Spillane, J. P., & Zeuli, J. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis, 21*(1), 1–27.

Spindler, G. (Ed.). (2000). *Fifty years of anthropology and education 1950–2000: A Spindler anthology.* Upper Saddle River, NJ: Lawrence Erlbaum.

Squire, J. R., & Applebee, R. K. (1968). *High school English instruction today: The national study of high school English programs.* New York: Irvington.

Stallings, J., & Kaskowitz, D. (1974). *Follow Through Classroom Observation evaluation 1972–73.* Menlo Park, CA: Stanford Research Institute.

Steele, J. M., House, E. R., & Kerins, T. (1971). An instrument for assessing instructional climate through low-inference student judgments. *American Educational Research Journal, 8,* 447–466.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*(1), 50–80.

Stein, M. K., Silver, E. A., & Smith, M. S. (1998). Mathematics reform and teacher development: A community of practice perspective. In J. Greeno & S. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 17–52). Mahwah, NJ: Lawrence Erlbaum.

Stein, M. K., Smith, M. S., & Silver, A. (1999). The development of professional developers: Learning to assist teachers in new settings in new ways. *Harvard Educational Review, 69*(3), 237–269.

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*(2), 87–100.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Stout, R. T. (1996). Staff development policy: Fuzzy choices in an imperfect market. *Education Policy Analysis Archives, 4*(2), 1–17.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions.* San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Supovitz, J. A. (2001). Translating teaching practice into improved student performance. In Fuhrman, S. (Ed.), *From the Capitol to the classroom: Standards-based reform in the states. 100th Yearbook of the National Society for the Study of Education, Part 2* (pp. 81–98). Chicago: University of Chicago Press.

Supovitz, J., & Turner, H. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*(9), 963–980.

Supovitz, J. A., & Zeif, S. G. (2000). Why they stay away. *Journal of Staff Development, 21*(4), 24–28.

Sykes, G. (1996). Reform of and as professional development. *Phi Delta Kappan, 77*(7), 465–489.

Talbert, J. E., & McLaughlin, M. W. (1993). Understanding teaching in context. In D. Cohen, M. McLaughlin, & J. Talbert (Eds.), *Teaching for understanding: Challenges for policy and practice* (pp. 167–206). San Francisco: Jossey-Bass.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Applied Social Research Methods No. 46). Thousand Oaks, CA: Sage.

Taylor, B. M., Pearson, P. D., Clark, K. F., & Walpole, S. (1999). *Beating the odds in teaching all children to read* (CIERA Report No. 2–006). Ann Arbor: Center for the Improvement of Early Reading Achievement, University of Michigan.

Teaching Commission. (2004). *Teaching at risk: A call to action.* New York: Author.

Thomas, G., Wineburg, S., Grossman, P., Myhre, O., & Woolworth, S. (1998). In the company of colleagues: An interim report on the development of a community of teacher learners. *Teaching and Teacher Education, 14,* 21–32.

Thompson, C. L., & Zeuli, J. S. (1999). The frame and the tapestry: Standards-based reform and professional development. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 341–375). San Francisco: Jossey-Bass.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.

Tourangeau, R., & Smith, T. (1998). Collecting sensitive information with different modes of data collection. In M. Couper, R. Baker, J. Bethlehem, C. Clark, J. Martin, W. Nicholls, & J. O'Reilly (Eds.), *Computer assisted survey information collection* (pp. 431–453). New York: John Wiley.

Turner, C. F., Ku, L. & Rogers, S. M. (1998). Adolescent sexual behaviour, drug use and violence: Increased reporting with computer survey technology. *Science, 280,* 867–873.

Ulewicz, M., & Beatty, A. (Eds.). (2001). *The power of video technology in international comparative research in education.* Board on Testing and Assessment Center for Education. Washington, DC: National Academy Press.

Von Secker, C. (2002). Effects of inquiry-based teacher practices on science excellence and equity. *Journal of Educational Research, 95,* 151–160.

Walberg, H., & Thomas, S. (1972). Open education: An operational definition and validation in Great Britain and the United States. *American Educational Research Journal, 9,* 197–207.

Walsh, W. (1967). Validity of self-report. *Journal of Counseling Psychology, 14*(1), 18–23.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher, 37*(8), 469–479.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge.* Los Altos, CA: Morgan Kaufmann.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10.* Retrieved September 18, 2008, from http://epaa.asu/edu/eppa/v10n12/

Wengraf, T. (2004). *Qualitative research interviewing.* Thousand Oaks, CA: Sage.

Wiley, D., & Yoon, B.(1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis, 17*(3), 355–370.

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. *Review of Research in Education, 24,* 173–209.

Wilson, S. M., Floden, R., & Ferrini-Mundy, J. (2002). Teacher preparation research: An insider's view from the outside. *Journal of Teacher Education, 53*(3), 190–204.

Wragg, E. C. (1999). *An introduction to classroom observations* (2nd ed.). New York: RoutledgeFalmer.

Wubbels, T., Brekelmans, M., & Hooymaters, H. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education, 8*(1), 47–58.

Yin, R., & Campbell, D. (2003). *Case study research: Design and methods* (Applied Social Science Research Methods Series, Vol. 5). Thousand Oaks, CA: Sage.

Yoon, K. S., Jacobson, R., Garet, M., Birman, B., & Ludwig, M. (2004, April). *Professional development activity log (PDAL): A new approach to design, measurement, data collection, and analysis.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

AUTHOR

LAURA M. DESIMONE is an associate professor of education policy at the University of Pennsylvania, Graduate School of Education, 3700 Walnut Street, Philadelphia, PA 19104; *lauramd@gse.upenn.edu.* Her research focuses on policy effects on teaching and learning, policy implementation, and the improvement of methods for studying policy effects and implementation (e.g., improving the quality of surveys and the appropriate use of multiple methodologies); much of her work is in the areas of standards-based reform/accountability and teacher quality initiatives (e.g., teachers' professional development, induction).

Manuscript submitted July 2, 2008
Revisions received October 3, 2008,
and November 26, 2008
Accepted December 1, 2008