ISSUES&ANSWERS REL 2007–No. 033

REL SOUTHWEST
Regional Educational Laboratory
At Edvance Research, Inc.

Reviewing the evidence on how teacher professional development affects student achievement

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

ISSUES & ANSWERS REL 2007–No. 033

REL SOUTHWEST
Regional Educational Laboratory
At Edvance Research, Inc.

Reviewing the evidence on how teacher professional development affects student achievement

October 2007

Prepared by

Kwang Suk Yoon
American Institutes for Research

Teresa Duncan
American Institutes for Research

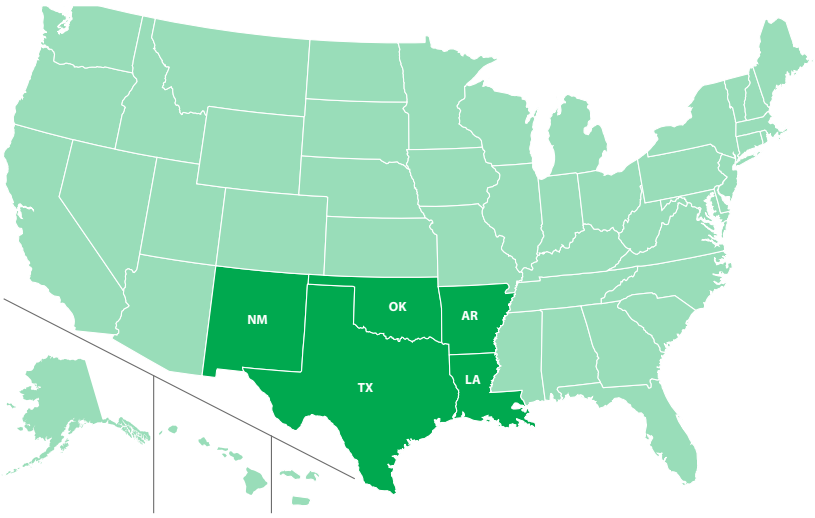
Silvia Wen-Yu Lee
American Institutes for Research

Beth Scarloss
American Institutes for Research

Kathy L. Shapley
Edvance Research

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL SOUTHWEST
At Edvance Research, Inc.



Issues & Answers is an ongoing series of reports from short-term Fast Response Projects conducted by the regional educational laboratories on current education issues of importance at local, state, and regional levels. Fast Response Project topics change to reflect new issues, as identified through lab outreach and requests for assistance from policymakers and educators at state and local levels and from communities, businesses, parents, families, and youth. All Issues & Answers reports meet Institute of Education Sciences standards for scientifically valid research.

October 2007

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-06-CO-0017 by Regional Educational Laboratory Southwest administered by Edvance Research. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs

This report is available on the regional educational laboratory web site at http://ies.ed.gov/ncee/edlabs.

Summary

Reviewing the evidence on how teacher professional development affects student achievement

Of the more than 1,300 studies identified as potentially addressing the effect of teacher professional development on student achievement in three key content areas, nine meet What Works Clearinghouse evidence standards, attesting to the paucity of rigorous studies that directly examine this link. This report finds that teachers who receive substantial professional development—an average of 49 hours in the nine studies—can boost their students' achievement by about 21 percentile points.

How does teacher professional development affect student achievement? The connection seems intuitive. But demonstrating it is difficult.

Examining more than 1,300 studies identified as potentially addressing the effect of teacher professional development on student achievement in three key content areas, this report finds nine that meet What Works Clearinghouse evidence standards. That only nine meet standards attests to the paucity of rigorous studies that directly assess the effect of in-service teacher professional development on student achievement in mathematics, science, and reading and English/language arts.

But the results of those studies—that average control group students would have increased their achievement by 21 percentile points if their teacher had received substantial professional development—indicates that providing professional development to teachers had a moderate effect on student achievement across the nine studies. The effect size was fairly consistent across the three content areas reviewed.

All nine studies focused on elementary school teachers and their students. About half focused on lower elementary grades (kindergarten and first grade), and about half on upper elementary grades (fourth and fifth grades).

Six studies were published in peer-reviewed journals; three were unpublished doctoral dissertations. The studies were not particularly recent, ranging from 1986 to 2003.

Five studies were randomized controlled trials that meet evidence standards without reservations. Four studies meet evidence standards with reservations (one randomized controlled trial with group equivalence problems and three quasi-experimental designs).

Four focused on student achievement in reading and English/language arts—unsurprising given the large literature in this content area. Two studies focused on mathematics, two on mathematics and reading and

English/language arts, one on science, and one on mathematics, science, and reading and English/language arts.

Only one effect of the 20 identified across the nine studies was negative, and only one effect was zero. The other 18 were positive. The sole negative effect was in a study of mathematics (fractions computation), where traditional instruction showed more positive effects on student achievement than a reform model. The effect was not statistically significant but was large enough to be considered substantively important. The sole zero effect was in a study of reading and English/language arts, where low-achieving students whose teachers were trained to use explicit instructional talk did not demonstrate appreciably greater reading achievement than their counterparts whose teachers attended a presentation on effective classroom management.

Studies that had more than 14 hours of professional development showed a positive and significant effect on student achievement from professional development. The three studies that involved the least amount of professional development (5–14 hours total) showed no statistically significant effects on student achievement.

All nine studies employed workshops or summer institutes. In all but one study follow-up sessions supported the main professional development event. The exception provided an intensive four-week summer workshop without follow-up support. In all nine studies professional development went directly to teachers rather than through a "train-the-trainer" approach and was delivered by the authors or their affiliated researchers.

Because of the lack of variability in form and the great variability in duration and intensity across the nine studies, discerning any pattern in these characteristics and their effects on student achievement is difficult. A larger number of rigorous studies on the link between professional development and student achievement might have made it possible to determine whether intensive, sustained, and content-focused professional development is more effective.

Highlighting the problems of many studies of professional development, this report can help researchers avoid methodological pitfalls. Especially important is that researchers undertaking studies with quasi-experimental designs provide data on the baseline equivalence of the treatment and comparison groups. Future studies of the effect of professional development on both teachers and students would be particularly useful—studies more fully addressing professional development's direct effect on teachers and its indirect effect on students.

October 2007

TABLE OF CONTENTS

Of the more than 1,300 studies identified as potentially addressing the effect of teacher professional development on student achievement in three key content areas, nine meet What Works Clearinghouse evidence standards, attesting to the paucity of rigorous studies that directly examine this link. The report finds that teachers who receive substantial professional development—an average of 49 hours in the nine studies—can boost their students' achievement by about 21 percentile points.

OVERVIEW

Professional development for teachers is a key mechanism for improving classroom instruction and student achievement (Ball & Cohen, 1999; Cohen & Hill, 2000; Corcoran, Shields, & Zucker, 1998; Darling-Hammond & McLaughlin, 1995; Elmore, 1997; Little, 1993; National Commission on Teaching and America's Future, 1996). Although calls for high quality professional development are perennial, there remains a shortage of such programs—characterized by coherence, active learning, sufficient duration, collective participation, a focus on content knowledge, and a reform rather than traditional approach (for details on one study of professional development, see box 1; for more information, see Garet, Porter, Desimone, Birman, & Yoon, 2001; Loucks-Horsley, Hewson, Love, & Stiles, 1998; National Commission on Teaching and America's Future, 1996; Birman et al., 2007; U.S. Department of Education, 2001).

A particular target for criticism is the prevalence of single-shot, one-day workshops that often make teacher professional development "intellectually superficial, disconnected from deep issues of curriculum and learning, fragmented, and noncumulative" (Ball & Cohen, 1999, pp. 3–4). And because there is no coherent infrastructure for professional development, professional development represents a "patchwork of opportunities—formal and informal, mandatory and voluntary, serendipitous and planned" (Wilson & Berne, 1999, p.174).

Recognizing the short supply of high quality professional development for teachers, the No Child Left Behind Act of 2001 mandated that teachers receive such learning opportunities. No Child Left Behind sets five criteria for professional development to be considered high quality:

- It is sustained, intensive, and content-focused—to have a positive and lasting impact on classroom instruction and teacher performance.

BOX 1

A study of professional development in mathematics

Birman et al. (2007) show that few teachers receive intensive, sustained, and content-focused professional development in mathematics. Teachers averaged 8.3 hours of professional development on how to teach mathematics and 5.2 hours on the "in-depth study" of topics in mathematics during the 12 months spanning the 2003/04 school year and the summer of 2004. Of elementary teachers, 71 percent participated in professional development focused on instructional strategies for teaching mathematics. But only 9 percent participated for more than 24 hours during the one-year period. Even fewer elementary school teachers (49 percent) reported that they participated in professional development focused on the in-depth study of mathematics during the same time period, and only 6 percent participated for more than 24 hours. Of secondary mathematics teachers, 51 percent attended professional development focused on the in-depth study of mathematics, but only 10 percent spent more than 24 hours on that content during the year.

- It is aligned with and directly related to state academic content standards, student achievement standards, and assessments.
- It improves and increases teachers' knowledge of the subjects they teach.
- It advances teachers' understanding of effective instructional strategies founded on scientifically based research.
- It is regularly evaluated for effects on teacher effectiveness and student achievement.

Because No Child Left Behind requires that activities supported by Title II funds be based on scientifically based research that shows how such interventions improve student achievement, better information on how professional development programs affect student achievement is an urgent need, both in the Southwest Region and nationally. This report reviews the research-based evidence on the effects of professional development on student achievement. The focus is on student achievement in three subjects: mathematics, science, and reading and English/language arts.

Examining more than 1,300 studies identified as potentially addressing the effect of teacher professional development on student achievement in the three subjects, this report identifies nine that meet What Works Clearinghouse evidence standards. That only nine meet standards attests to the paucity of rigorous studies that directly examine the effect of in-service teacher professional development on student achievement.

But the results of those studies—that average control group students would have increased their achievement by 21 percentile points if their teacher had received substantial professional development—indicates that providing professional development to teachers had a moderate effect on student achievement across the nine studies. The effect size was fairly consistent across the three content areas reviewed.

All nine studies focused on elementary school teachers and their students. About half focused on lower elementary grades (kindergarten and first grade), and about half on upper elementary grades (fourth and fifth grades).

Six studies were published in peer-reviewed journals; three were unpublished doctoral dissertations. The studies were not particularly recent, ranging from 1986 to 2003.

Five studies were randomized controlled trials that meet evidence standards without reservations. Four studies meet evidence standards with reservations (one randomized controlled trial with group equivalence problems and three quasi-experimental designs).

Four focused on student achievement in reading and English/language arts—unsurprising given the large literature in this content area. Two

studies focused on mathematics, two on mathematics and reading and English/language arts, one on science, and one on mathematics, science, and reading and English/language arts.

Only one effect of the 20 identified across the nine studies was negative, and only one effect was zero. The other 18 were positive. The sole negative effect was in a study of mathematics (fractions computation), where traditional instruction showed more positive effects on student achievement than a reform model. The effect was not statistically significant but was large enough to be considered substantively important. The sole zero effect was in a study of reading and English/language arts, where low-achieving students whose teachers were trained to use explicit instructional talk did not demonstrate appreciably greater reading achievement than their counterparts whose teachers attended a presentation on effective classroom management.

Studies that had more than 14 hours of professional development showed a positive and significant effect on student achievement from professional development. The three studies that involved the least amount of professional development (5–14 hours total) showed no statistically significant effects on student achievement.

All nine studies employed workshops or summer institutes. In all but one study follow-up sessions supported the main professional development event. The exception provided an intensive four-week summer workshop without follow-up support. In all nine studies professional development went directly to teachers rather than through a "train-the-trainer" approach and was delivered by the authors or their affiliated researchers.

Because of the lack of variability in form and the great variability in duration and intensity across the nine studies, discerning any pattern in these characteristics and their effects on student achievement is difficult. A larger number of rigorous studies on the link between professional development and student achievement might have made it possible to determine whether intensive, sustained, and content-focused professional development is more effective.

Highlighting the problems of many studies of professional development, this report can help researchers avoid methodological pitfalls. Especially important is that researchers undertaking studies with quasi-experimental designs provide data on the baseline equivalence of the treatment and comparison groups. Future studies of the effect of professional development on both teachers and students would be particularly useful—more fully addressing professional development's direct effect on teachers and its indirect effect on students.

DEMONSTRATING THE EFFECT OF TEACHER PROFESSIONAL DEVELOPMENT ON STUDENT ACHIEVEMENT

Showing that professional development translates into gains in student achievement poses tremendous challenges, despite an intuitive and logical connection (Borko, 2004; Loucks-Horsley & Matsumoto, 1999; Supovitz, 2001. To substantiate the empirical link between professional development and student achievement, studies should ideally establish two points. One is that there are links among professional development, teacher learning and practice, and student learning. The other is that the empirical evidence is of high quality—that the study proves what it claims to prove. This report focuses on the second point, treating the first only briefly.

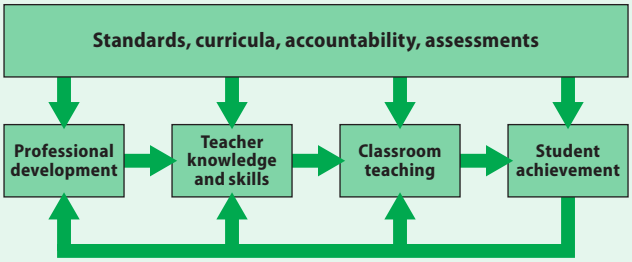
Showing that professional development translates into gains in student achievement poses tremendous challenges, despite an intuitive and logical connection

The links among professional development, teacher learning and practice, and student achievement

Consistent with models of effective professional development (Cohen & Hill, 2000; Fishman, Marx, Best, & Tal, 2003; Garet et al., 2001; Guskey

& Sparks, 2004; Kennedy, 1998; Loucks-Horsley & Matsumoto, 1999), this report assumes that professional development's effects on student achievement are mediated by teacher knowledge and practice in the classroom and that professional development takes place in the context of high standards, challenging curricula, system-wide accountability, and high-stakes assessments (figure 1).

FIGURE 1
How professional development affects student achievement



Professional development affects student achievement through three steps. First, professional development enhances teacher knowledge and skills. Second, better knowledge and skills improve classroom teaching. Third, improved teaching raises student achievement. If one link is weak or missing, better student learning cannot be expected. If a teacher fails to apply new ideas from professional development to classroom instruction, for example, students will not benefit from the teacher's professional development.

In the first step, professional development must be of high quality in its theory of action, planning, design, and implementation.

- It should be intensive, sustained, content-focused, coherent, well defined, and strongly implemented (Garet et al., 2001; Guskey, 2003; Loucks-Horsley, Hewson, Love, & Stiles, 1998; Supovitz, 2001; Wilson & Berne, 1999).
- It should be based on a carefully constructed and empirically validated theory of teacher learning and change (Ball & Cohen, 1999; Richardson & Placier, 2001; Sprinthall, Reiman, & Thies-Sprinthall, 1996).
- It should promote and extend effective curricula and instructional models—or materials based on a well defined and valid theory of action (Cohen, Raudenbush, & Ball, 2002; Hiebert & Grouws, 2007; Rossi, Lipsey, & Freeman, 2004).

In the second step, teachers must have the motivation, belief, and skills to apply the professional development to classroom teaching (Borko, 2004; Showers, Joyce, & Bennett, 1987), supported by ongoing school collaboration and follow-up consultations with experts. Doing so could require overcoming such barriers to new practices as lack of time for preparation and instruction, limited materials and human resources, and lack of follow-up support from professional development providers.

In the third step, teaching—improved by professional development—raises student achievement. The challenge is evaluating the gains.

The quality of empirical evidence

Establishing the second point—that the empirical evidence is of high quality—is the primary focus of this report, which examines the rigor of empirical studies conducted to validate the effects of professional development (National Research Council, 2004). Even if professional development enhances teacher knowledge and skills and improves classroom instruction, a poorly designed evaluation or inadequate implementation would make it difficult to detect any effects from the professional development.

What is required for establishing the empirical link between professional development and student achievement? That empirical link is based on at least four elements:

- A rigorous research design must ensure the internal validity of causal inferences about

the effectiveness of professional development. Using a study design with strong internal validity (a randomized controlled trial, for example) can rule out competing explanations for gains in student academic achievement. The research design should be able to measure the value that professional development adds to student learning separately from the value added by innovative curricula, instruction, or materials. A rigorous research design must also have externally valid findings, adequate statistical power to detect true effects, and sufficient time between the professional development and the measurement of teacher and student outcomes.

- The study design must be executed with high fidelity and sufficient implementation of professional development

- Psychometric properties of measures must be adequate (measures of classroom teaching practices, of student achievement, and of teacher knowledge, beliefs, and behaviors). Measures should be valid, reliable, age-appropriate, and sensitive to and aligned with the intervention.

- Analytic models must be well-specified and statistical methods must be appropriate

Given these requirements, it is unsurprising that few rigorous studies address the effect of professional development on student achievement (Borko, 2004; Clewell, Campbell, & Perlman, 2004; Kennedy, 1998; Killion, 1999; Loucks-Horsley & Matsumoto, 1999; Supovitz, 2001). There is more literature on the effects of professional development on teacher learning and teaching practice, falling short of demonstrating effects on student achievement (Garet et al., 2001). In addition, even more literature addresses curricular or instructional effectiveness (National Research Council, 2004; various What Works Clearinghouse intervention reports).

One systematic review of the effects of professional development on student achievement is Kennedy (1998). That review analyzes the relative effects on student outcomes from professional development programs for math and science, examining the professional development's subject, content focus, skill level, form, and other features (intensity and concentration, for example). The conclusion:

Few rigorous studies address the effect of professional development on student achievement—there is more literature on the effects of professional development on teacher learning and teaching practice

> Programs whose content focused mainly on teachers' behaviors demonstrated smaller influences on student learning than did programs whose content focused on teachers' knowledge of the subject, on the curriculum, or on how students learn the subject (p. 18).

Kennedy's seminal review indicates the importance of content focus in high quality professional development (see also Desimone, Porter, Garet, Yoon, & Birman, 2002; Garet et al., 2001; Yoon, Garet, Birman, & Jacobson, 2007). There are three reasons, however, for a new systematic review to supplement those of Kennedy and of Clewell, Campbell, and Perlman (2004).

First, the volume of literature has grown, especially after standards-based reform prompted a wave of professional development–related studies. Second, most of the literature reviews and research syntheses are limited in scope, source, and subject. Few literature reviews encompass the three core academic subjects under No Child Left Behind accountability requirements (mathematics, science, and reading and English/language arts). A more comprehensive and systematic review of evidence that professional development works in these critical subject areas is needed. Third, the growing emphasis on effective professional development practices supported by scientifically based research makes it imperative to apply rigorous evidence standards—such as those of the What Works Clearinghouse—in new literature reviews and syntheses.

NINE STUDIES THAT MEET EVIDENCE STANDARDS

This report reviewed more than 1,300 studies to identify those that potentially addressed the impact of teacher professional development on student achievement. Only nine meet What Works Clearinghouse evidence standards—attesting to the paucity of rigorous studies that directly examine the effect of in-service teacher professional development on student achievement in the three core academic subjects. For studies not meeting evidence standards despite focusing on teacher professional development and including a student achievement measure, a frequent problem was study design, particularly for quasi-experimental designs with problems in baseline equivalence between treatment and comparison groups (for details on the methodology and the studies that did not meet evidence standards, see box 2 and appendix A). The nine studies:

- Carpenter, Fennema, Peterson, Chiang, & Loef (1989).
- Cole (1992).
- Duffy et al. (1986).
- Marek & Methven (1991).
- McCutchen et al. (2002).
- McGill-Franzen, Allington, Yokoi, & Brooks (1999).
- Saxe, Gearhart, & Nasir (2001).
- Sloan (1993).
- Tienken (2003).

For studies not meeting evidence standards, a frequent problem was study design, particularly for quasi-experimental designs with problems in baseline equivalence between treatment and comparison groups

All nine studies focused on elementary school teachers and their students. About half focused on lower elementary grades (kindergarten and first grade), and about half on upper elementary grades (fourth and fifth grades).

Six studies were published in peer-reviewed journals; three were unpublished doctoral dissertations. The studies were not particularly recent, ranging from 1986 to 2003.

Five studies were randomized controlled trials that meet evidence standards without reservations. Four studies meet evidence standards with reservations (one randomized controlled trial with group equivalence problems and three quasi-experimental designs).

Four focused on student achievement in reading and English/language arts—unsurprising given the large literature in this content area. Two studies focused on mathematics, two on mathematics and reading and English/language arts, one on science, and one on mathematics, science, and reading and English/language arts.

Seven studies used standardized measures of achievement. One used researcher-developed measures of students' knowledge of fractions, and one used Piagetian conservation tasks as the outcome.

Studies were usually of teachers and their intact classrooms. Two studies randomly sampled students from each teacher's classroom, and one focused only on low-achieving readers in the classrooms. The number of teachers ranged from 5 in one study to 44 in another, with student sample sizes ranging from 98 to 779.[1] Clustering of students within classrooms was typically not addressed in the studies. This report therefore applies clustering corrections to the reported statistical significance of the findings. When necessary, corrections are also applied for multiple outcomes to decrease the familywise error rates.

EFFECTS OF PROFESSIONAL DEVELOPMENT ON STUDENT ACHIEVEMENT IN THE NINE STUDIES

Twenty effect sizes and improvement indices were computed across the nine studies (table 1; see box 2 for methodology and definitions).

- The average effect size across the nine studies was 0.54, ranging from –0.53 to 2.39.

BOX 2

Methodology

Understanding how this report reviewed the research-based evidence on the effectiveness of professional development is important background for interpreting the results.

Review protocol

Developing a review protocol was the first step in systematically reviewing the research-based evidence on the effectiveness of professional development. The approach was modeled on the review process and rigorous evidence standards of the U.S. Department of Education's What Works Clearinghouse. The protocol established the relevance criteria for literature searches and the parameters for screening and reviewing studies (see appendix B for the full protocol). Criteria included:

- *Topic*. The study had to deal with the effects of in-service teacher professional development on student achievement.
- *Population*. The sample had to include teachers of English, mathematics, and science and their students in grades K–12.
- *Study design*. The review of evidence was limited to final manuscripts that were based on empirical studies using randomized controlled trials or quasi-experimental designs. In randomized controlled trials participants are randomly assigned to different experimental groups. Quasi-experimental designs do not randomly assign participants to intervention and comparison groups, but the groups are matched or shown to be equivalent before the intervention.
- *Outcome*. The study had to measure student achievement outcomes.
- *Outcome measure validity*. The study had to use measures demonstrated to be accurate and consistent.
- *Time*. The study had to be conducted between 1986 and 2006.
- *Country*. Studies had to take place in Australia, Canada, the United Kingdom, or the United States—due to concerns about the external validity of the findings.

Studies were then gathered through an extensive electronic search of published and unpublished research. Fourteen key researchers were also asked to identify studies. Eight researchers responded, recommending additional studies that fit the study purpose. Submitted to the prescreening process were 1,343 studies. Of these, 907 were unique. The remaining 436 studies were duplicates but were included in the final tally because they addressed multiple subject areas (math and science, for example).

Screening and coding

Screening and coding were conducted by six doctorate-level analysts over four months—in four stages: prescreening, stage 1–full screening, stage 2–coding, and stage 3–coding.

Only 132 unique studies met all five criteria in the prescreening step and went to the next stage of review. The 27 studies that passed the stage 1 full screening were subject to stage 2 coding. Only nine studies met evidence standards and were submitted to the final stage of coding.

The nine studies that "met evidence standards" or "met evidence standards with reservations" were reviewed to describe important characteristics of the study and the professional development. These characteristics included:

- Estimated impact of the professional development (in effect sizes and improvement indices).
- Replicability of the professional development and the study.
- Teacher outcome measures.
- Content, form, and other features of the professional development (using the classification in Kennedy, 1998).
- Whether the effect of professional development was confounded with that of curriculum.
- Statistical analysis.
- Statistical reporting.

Effect sizes and improvement indices

Effect sizes and improvement indices were computed using the formulas of the What Works Clearinghouse.[1] An effect size—a standardized mean

(CONTINUED)

BOX 2 (CONTINUED)

Methodology

difference—expresses in standard deviation units the increase or decrease in achievement of the intervention group compared with that of the control or comparison group. The improvement index is the difference between the percentile rank corresponding to the intervention group mean and the percentile rank corresponding to the control group mean in the control group distribution (the 50th percentile). So, the improvement index can be interpreted as the expected change in percentile rank for an average control group student if the student had received the intervention (for the studies in this report, if the student was in a classroom with a teacher who had received professional development).

Statistical significance and substantive importance

Consistent with What Works Clearinghouse procedures, the statistical significance of the effect sizes was corrected as necessary to adjust for unaccounted clustering and for multiple outcomes. Effect sizes whose absolute values were 0.25 or greater are labeled "substantively important." The results in this report are overall results for the student samples rather than effects by subgroup (those analyses are also highly underpowered). The only exception is McCutchen et al. (2002), where an effect size could be computed only for the kindergarten subsample.

Note

1. Details are available from the What Works Clearinghouse web site (http://www.whatworks.ed.gov/reviewprocess/conducted_computations.pdf).

- The average improvement index was 21, ranging from –20 to 49.
- Only one effect was negative (in Saxe et al., 2001), and only one effect was zero (in Duffy et al., 1986). The other 18 effects were positive, with effect sizes ranging from 0.12 to 2.39 (with improvement indices from 5 to 49).
- Of the 20 effects, 12 were not statistically significant after applying necessary corrections for unaddressed clustering and multiple outcomes. Nine of those twelve, however, are substantively important according to What Works Clearinghouse conventions.
- Fifteen of the effects came from the five randomized controlled trials that meet What Works Clearinghouse standards. The average effect size for the randomized controlled trials was 0.51, ranging from 0 to 1.11.
- Five of the effects came from four studies that meet What Works Clearinghouse standards with reservations (three quasi-experimental designs and one problematic randomized controlled trial). The average effect size was 0.61, ranging from –0.53 to 2.39.

Effects by content area

Disaggregating the studies by their content-area outcomes allowed computing averages and ranges for science, mathematics, and reading and English/language arts (table 2). Science had only 2 effects, mathematics had 6, and reading and English/language arts had 12. The average effect was remarkably consistent across the three content areas. The average effect size in science was 0.51; in mathematics, 0.57; and in reading and English/language arts, 0.53.

The sole negative effect (with an effect size of –0.53, in Saxe et al., 2001) was in mathematics (fractions computation), where traditional instruction showed more positive effects on student achievement than a reform model. The effect was not statistically significant but was large enough to be substantively important. The sole zero effect was in reading and English/language arts (in Duffy et al., 1986), where low-achieving students whose teachers were trained to use explicit instructional talk did not demonstrate appreciably greater reading achievement than their counterparts whose teachers attended a presentation on effective classroom management.

TABLE 1

Effects of professional development on student achievement, by study

Study (study design)	Outcome measure	Effect size	Applied correction for clustering or multiple comparisons?	Recomputed statistical significance	Improvement index
Carpenter et al., 1989 (RCT)	Iowa Test of Basic Skills Level 7, computation	0.41	None applied if author did not report significant results	Not significant, but substantively important	16
	Iowa Test of Basic Skills Level 7, problem solving	0.41	None applied if author did not report significant results	Not significant, but substantively important	16
Cole, 1992 (RCT)	Average for math	0.50	Yes	Statistically significant	19
	Average for reading	0.82	Yes	Statistically significant	29
	Average for language	0.24	None applied if author did not report significant results	Not significant	9
Duffy et al., 1986 (RCT)	Gates-MacGinitie Reading Test	0.00	None applied if author did not report significant results	Not significant	0
Marek & Methven, 1991 (QED)	Average for conservation test	0.39	Yes	Statistically significant	15
McCutchen et al., 2002 (QED)	Gates-MacGinitie Word Reading Subtest	0.39	No	Statistically significant	15
McGill-Franzen et al., 1999 (RCT)	Concepts about print	1.11	Yes	Statistically significant	37
	Letter identification	0.69	Yes	Statistically significant	25
	Writing vocabulary	0.32	Yes	Not significant, but substantively important	13
	Ohio Word Test	0.66	Yes	Not significant, but substantively important	24
	Hearing sounds in words	0.97	Yes	Statistically significant	33
	Peabody Picture Vocabulary Test	0.12	None applied if author did not report significant results	Not significant	5
Saxe et al., 2001 (QED)	Fraction concepts	2.39	No	Statistically significant	49
	Fraction computation	–0.53	No	Not significant, but substantively important	–20
Sloan, 1993 (RCT)	Comprehensive Test of Basic Skills, reading	0.68	Yes	Not significant, but substantively important	25
	Comprehensive Test of Basic Skills, mathematics	0.26	None applied if author did not report significant results	Not significant, but substantively important	10
	Comprehensive Test of Basic Skills, science	0.63	Yes	Not significant, but substantively important	23
Tienken, 2003 (RCT with group equivalence problems)	Content/organization score on narrative writing test	0.41	Yes	Not significant, but substantively important	16
	Average effect size across all studies	*0.54*		*Average improvement index across all studies*	*21*
	Minimum effect size across all studies	*–0.53*		*Minimum improvement index across all studies*	*–20*
	Maximum effect size across all studies	*2.39*		*Maximum improvement index across all studies*	*49*

RCT is a randomized controlled trial; QED is a quasi-experimental design.

Source: Authors' calculations based on data described in text.

TABLE 2

Effects of professional development on student achievement, by content area

Study (study design)	Outcome measure	Effect size	Applied correction for clustering or multiple comparisons?	Recomputed statistical significance	Improvement index
Science					
Marek & Methven, 1991 (QED)	Average for conservation test	0.39	Yes	Statistically significant	15
Sloan, 1993 (RCT)	Comprehensive Test of Basic Skills, science	0.63	Yes	Not significant, but substantively important	23
	Content area average effect size	*0.51*		*Content area average improvement*	*19*
	Content area minimum effect size	*0.39*		*Content area minimum improvement*	*15*
	Content area maximum effect size	*0.63*		*Content area maximum improvement*	*23*
Mathematics					
Carpenter et al., 1989 (RCT)	Iowa Test of Basic Skills Level 7, computation	0.41	None applied if author did not report significant results	Not significant, but substantively important	16
Carpenter et al., 1989 (RCT)	Iowa Test of Basic Skills Level 7, problem solving	0.41	None applied if author did not report significant results	Not significant, but substantively important	16
Cole, 1992 (RCT)	Average for math	0.50	Yes	Statistically significant	19
Saxe et al., 2001 (QED)	Fraction concepts	2.39	No	Statistically significant	49
Saxe et al., 2001 (QED)	Fraction computation	–0.53	No	Not significant, but substantively important	–20
Sloan, 1993 (RCT)	Comprehensive Test of Basic Skills, mathematics	0.26	None applied if author did not report significant results	Not significant, but substantively important	10
	Content area average effect size	*0.57*		*Content area average improvement*	*22*
	Content area minimum effect size	*–0.53*		*Content area minimum improvement*	*–20*
	Content area maximum effect size	*2.39*		*Content area maximum improvement*	*49*

Study (study design)	Outcome measure	Effect size	Applied correction for clustering or multiple comparisons?	Recomputed statistical significance	Improvement index
Reading and English/language arts					
Cole, 1992 (RCT)	Average for reading	0.82	Yes	Statistically significant	29
Cole, 1992 (RCT)	Average for language	0.24	None applied if author did not report significant results	Not significant	9
Duffy et al., 1986 (RCT)	Gates-MacGinitie Reading Test	0.00	None applied if author did not report significant results	Not significant	0
McCutchen et al., 2002 (QED)	Gates-MacGinitie Word Reading Subtest (Kindergarten sample)	0.39	No	Statistically significant	15
McGill-Franzen et al., 1999 (RCT)	Concepts about print	1.11	Yes	Statistically significant	37
McGill-Franzen et al., 1999 (RCT)	Letter identification	0.69	Yes	Statistically significant	25
McGill-Franzen et al., 1999 (RCT)	Writing vocabulary	0.32	Yes	Not significant, but substantively important	13
McGill-Franzen et al., 1999 (RCT)	Ohio Word Test	0.66	Yes	Not significant, but substantively important	24
McGill-Franzen et al., 1999 (RCT)	Hearing sounds in words	0.97	Yes	Statistically significant	33
McGill-Franzen et al., 1999 (RCT)	Peabody Picture Vocabulary Test	0.12	None applied if author did not report significant results	Not significant	5
Sloan, 1993 (RCT)	Comprehensive Test of Basic Skills, reading	0.68	Yes	Not significant, but substantively important	25
Tienken, 2003 (RCT with group equivalence problems)	Content/organization on narrative writing task	0.41	Yes	Not significant, but substantively important	16
	Content area average effect size	*0.53*		*Content area average improvement*	*20*
	Content area minimum effect size	*0.00*		*Content area minimum improvement*	*0*
	Content area maximum effect size	*1.11*		*Content area maximum improvement*	*37*

RCT is a randomized controlled trial; QED is a quasi-experimental design.

Source: Authors' calculations based on data described in text.

BOX 3

Kennedy's professional development content groups

Kennedy's (1998) classification scheme for professional development differentiates between four types.

Group 1 focused on teaching behaviors applying generically to all subjects. These behaviors might result from process-product research or might include strategies such as cooperative grouping. The methods are expected to be equally effective across school subjects.

Group 2 focused on teaching behaviors applying to a particular subject. Although presented for a particular subject, the behaviors have a generic quality and are expected to be generally applicable in that subject.

Group 3 focused on curriculum and pedagogy, justified by how students learn. Such professional development provides general guidance on curriculum and pedagogy for teaching a subject and justifies its recommendations using knowledge about how students learn the subject.

Group 4 focused on how students learn and how to assess student learning. Such professional development provides knowledge about how students learn particular subjects but does not provide specific guidance on practices for teaching the subject.

Effects by form, contact hours, intensity, and duration of professional development

All nine studies employed workshops or summer institutes. In all but one follow-up sessions supported the main professional development event (see table 3 on page 15). Marek and Methven (1991) was the exception; that study provided an intensive four-week summer workshop without follow-up support. In all nine studies professional development went directly to teachers rather than through a train-the-trainer approach and was delivered by the authors or their affiliated researchers.

The professional development in these studies varied in duration and intensity. The total contact hours ranged from 5 hours to 100. Marek and Methven (1991) provided 100 hours of professional development over four weeks, while McCutchen et al. (2002) provided about the same number of contact hours but over 10 months, offering more sustained, if less intensive, development. Studies that had greater than 14 hours of professional development showed a positive and significant effect on student achievement from professional development. The three studies that involved the least amount of professional development (5–14 hours total) showed no statistically significant effects on student achievement.

Because of the lack of variability in form and the great variability in duration and intensity in this small number of studies, discerning any pattern between these characteristics and their effects on student achievement is difficult. A larger number of rigorous studies on the link between professional development and student achievement might have made it possible to determine whether intensive, sustained, and content-focused professional development is more effective (Ball & Cohen, 1999; Garet et al., 2001; Joyce & Showers, 1995; Loucks-Horsley, Stiles, & Hewson, 1996; Wilson & Berne, 1999; Yoon et al., 2007).

Effects by models and theories of action of professional development

The fourfold content-group classification scheme for professional development in Kennedy (1998) helps characterize the professional development models and theories of actions in the nine studies (box 3). The professional development in the nine studies varied much more in content and substance than in form—as predicted in Kennedy (1998). Likewise, Spillane (2000, p. 23) notes that "structural similarities in district professional development approaches (e.g., classroom demonstrations, peer coaching) camouflaged substantial differences in the underlying theories of teacher

learning and change." The limited number of studies and the variability in their professional development models precludes drawing any definitive links between content-group classification and effects on student achievement. Even so, a qualitative summary of the professional development approaches in the nine studies is a useful first step.

Cole (1992) and Sloan (1993) used a similar professional development model, focused on changes in teachers' behaviors applying generically to all subjects (group 1 in Kennedy's classification; see box 3 for details). In Cole (1992) teachers were trained to model 14 pedagogical behavior competencies—expected to apply generically to all subjects—specified in the Mississippi Teacher Assessment Instrument. In Sloan (1993) teachers practiced instructional and questioning behaviors recommended by the Direct Instruction model. Both studies tested the effects of this prescriptive and generic professional development on student achievement in multiple subjects by using commercial tests such as the Comprehensive Test of Basic Skills and the Stanford Achievement Test. Although all the effects were positive and favored the treatment group, none was statistically significant after adjusting for clustering and multiple outcomes. Five effects were large enough, however, to be considered substantively important (see tables 1 and 2).

In Duffy et al. (1986) teachers participated in professional development that focused on using explicit verbal explanations during reading instruction to poor readers (group 2 in Kennedy's classification, characterized by prescriptive, content-specific approaches that focus on changing teachers' behaviors). The study found no appreciable increase reading achievement.

In Marek and Methven (1991) teachers attended a workshop focused on science as knowledge and knowledge-seeking. The goal was to develop a curriculum of learning cycles representing this philosophy. McGill-Franzen et al. (1999) trained teachers to structure their classrooms and instruction to meet their young students' needs in literacy development. In Tienken (2003) teachers were trained to teach students to use a writing scoring rubric and high-order reflective questions as self-assessment devices in narrative writing. Common across these disparate professional development activities is a focus on curriculum or pedagogy justified by how students learn—group 3 of Kennedy's classification.

The professional development in the nine studies varied much more in content and substance than in form

Marek and Methven (1991) found statistically significant effects from professional development on students' conservation reasoning (as measured by Piagetian tasks). Although all six effect sizes in McGill-Franzen et al. (1999) were positive, only three were statistically significant after adjusting for clustering and multiple outcomes. Two of the three effects that were not statistically significant were large enough to be considered substantively important. The professional development in Tienken (2003) also had a substantively important—but not statistically significant—positive effect on students' narrative writing, after applying a clustering correction.

Carpenter et al. (1989) and Saxe et al. (2001) focused on increasing teachers' knowledge of students' mathematical thinking. McCutchen and et al. (2002) tried to boost teachers' knowledge of phonology and its link to orthography. Carpenter et al. (1989) and McCutchen et al. (2002) found positive effects on student achievement of about 0.40 (substantively important but not statistically significant in Carpenter et al. and statistically significant in McCutchen et al.). Saxe et al. (2001) found mixed effects. Large, positive, statistically significant effects on students' conceptual understanding of fractions favored the reform model. But negative and substantively important, though not statistically significant, effects on students' fraction computation skills in the reform model favored traditional instruction. These three professional development approaches allowed more teacher discretion in classroom teaching, focusing on deepening teachers' content knowledge and understanding of how students learn—group 4 of Kennedy's classification.

BETTER EVALUATION FOR BETTER PROFESSIONAL DEVELOPMENT

Few studies meet evidence standards. But the average effect size of 0.54 in mathematics, science, and reading and English/language arts—and the consistency of that effect size—indicates that providing professional development to teachers has a moderate effect on student achievement across the nine studies. Average control group students would have increased their achievement by 21 percentile points if their teacher had received professional development.

Average control group students would have increased their achievement by 21 percentile points if their teacher had received professional development

Results in mathematics are of particular note, given the data on professional development in mathematics in Birman et al. (2007; see box 1 for details). Four studies in mathematics reviewed here generated six effects, averaging 0.57, with an improvement index of 22 percentile points. The contact hours in the four studies averaged just over 53 hours, ranging from 30 hours to 83 hours, over a period of four months to one year. This professional development is longer than that of the typical elementary school teacher—only 9 percent of elementary school teachers participated in mathematics professional development for more than 24 hours over a year in Birman et al. (2007).

This report cannot determine definitively whether the professional development in the four studies meets other criteria for high quality professional development in the literature (using active learning and collective participation, for example) or in No Child Left Behind (consistent with state academic content standards and involving strategies from scientifically based research). Even so, the gap between the amount of professional development found effective in the four studies and the average received by elementary school teachers is worth considering.

These findings are important, but note four caveats:

First, none of the nine studies focused on professional development's effects on middle or high school students.

Second, even the studies meeting evidence standards were generally underpowered and did not address clustering or multiple comparisons. As a result, 12 effects of 20 were not statistically significant. The limited number of studies and the variability in their professional development approaches preclude any conclusions about the effectiveness of specific professional development programs or about the effectiveness of professional development by form, content, or intensity. Greater resources and time would allow a more comprehensive literature search for comparison. Using different keywords for search might generate a larger pool, for example. And more studies might meet evidence standards if authors could be contacted for additional information.

Third, each of the 9 studies and the 20 effects are treated equally, regardless of differences in type of professional development, sample sizes, or quality of research design. Because some studies included several outcome measures, those studies are overrepresented in the average overall effect. For example, McGill-Franzen et al. (1999) accounts for six effect sizes in the overall average.

Fourth, the report conducts none of the additional data manipulations of traditional meta-analysis, such as differential weighting. The intent was to adhere as closely as possible to What Works Clearinghouse procedures.[2] Although the What Works Clearinghouse computes an average effect size for a study and uses an average of study averages to report an overall average effect size, the studies in a What Works Clearinghouse intervention report address one intervention. This report, however, addresses several interventions, and the studies were few enough to merit limiting any additional aggregation, given the diversity among the nine studies in content areas and professional development approaches. So, the individual effects and the overall average are the only ones included. Interpreting the overall average effect size of 0.54

TABLE 3

Features of professional development in the nine studies that meet evidence standards

Study (study design)	Name or type of professional development	Philosophy	Content	Provider and delivery	Contact hours and duration	Kennedy content group[a]
Carpenter et al., 1989 (randomized controlled trial)	Cognitively guided instruction	Giving teachers knowledge from research on students' thinking and learning about mathematics changes teachers' teaching. It also improves how teachers assess student knowledge, which in turn changes teachers' instruction.	How students learn math, relationships between math problems and how students process to solve them, research on math acquisition, examination of curricula, how materials affect teaching, planning instruction	Authors Four-week workshop and one follow-up meeting	83 hours over four months	Group 4: Focused on how students learn and how to assess student learning
Cole, 1992 (randomized controlled trial)	Mississippi Teacher Assessment Instrument staff development	Teachers who use state-defined competencies will teach better, and therefore their students will learn more.	Modeling of the 14 Mississippi Teacher Assessment Instrument teacher (pedagogical) behavior competencies (for example, planning instruction to achieve selected objectives, organizing instruction to take into account individual differences among learners, and obtaining and using information about the needs and progress of individual learners)	Mississippi State Department of Education Eight three-hour sessions over a two-month period with follow-up observational visits throughout the year, plus two half day follow-up conferences	40+ hours over a year	Group 1: Focused on teaching behaviors applying generically to all subjects
Duffy et al., 1986 (randomized controlled trial)	Incorporating explicit verbal explanations during reading instruction	Training teachers in the use of explicit verbal explanations during reading instruction to poor readers will increase student awareness of what was taught, which in turn will enhance students' strategic reading skills.	How to recast teacher skill at prescriptive basal text techniques into strategies for helping students be better readers when removing blockages to meanings; how to make explicit statements about the reading skills being taught; how to organize these statements for presentation to students	Authors using Houghton Mifflin basal text Five two-hour sessions	10 hours over four months	Group 2: Focused on teaching behaviors applying to a particular subject
Marek & Methven, 1991 (quasi-experimental design)	Utilizing the Learning Cycle in Elementary School Science	Teaching science as a search for knowledge will lead students to construct their own knowledge about the world around them.	How to develop a curriculum (learning cycles) that represents science, allows students to experience science as a search for knowledge, and is compatible with their students' learning abilities.	National Science Foundation–funded workshop delivered by the Science Education Center at the University of Okalahoma Four-week summer workshop	100 hours over four weeks	Group 3: Focused on curriculum and pedagogy, justified by how students learn

(CONTINUED)

TABLE 3 (CONTINUED)

Features of professional development in the nine studies that meet evidence standards

Study (study design)	Name or type of professional development	Philosophy	Content	Provider and delivery	Contact hours and duration	Kennedy content group[a]
McCutchen et al., 2002 (quasi-experimental design)	n/a	Teachers should incorporate phonological awareness instruction into classroom practices. A deep understanding of phonology, pronunciation, reading skill development, and the links among them must be used in the classroom.	Deepening teachers' understanding of phonology, phonological awareness, analysis of sounds, development of phonological awareness in children, children's mistakes revealing underlying conception of phonemics	University research team Two-week summer institute plus three follow-up meetings; informal interactions and classroom visits with support	About 100 hours over 10 months	Group 4: Focused on how students learn and how to assess student learning
McGill-Franzen et al., 1999 (randomized controlled trial)	n/a	Improving children's access to books in their classrooms is not enough to develop literacy among kindergartners. It must be supplemented by enhancing their teachers' instructional routines involving the book collection.	Physical design of the classroom; effective book displays; importance of reading aloud to children; environmental print; author, genre, and content themes created with the book collection; small-group lessons using teacher-made materials based on books read	Authors Three whole-day sessions and seven two-hour follow-up sessions	About 30 hours over six months	Group 3: Focused on curriculum and pedagogy, justified by how students learn
Saxe et al., 2001 (quasi-experimental design)	n/a	Although good curriculum materials can provide rich tasks and activities that support students' mathematical investigations, such materials may not be sufficient to enable deep changes in instructional practice. With professional development, teachers must transform the ways they use curriculum materials with students.	Teacher knowledge of mathematics (particularly fractions), teacher knowledge of how students learn mathematics and fractions, and teacher understanding of student motivation in math	Authors/university-based developer A weeklong summer workshop with 13 follow-up meetings	About 60 hours over six and a half months	Group 4: Focused on how students learn and how to assess student learning

Study (study design)	Name or type of professional development	Philosophy	Content	Provider and delivery	Contact hours and duration	Kennedy content group[a]
Sloan, 1993 (randomized controlled trial)	n/a	Training teachers to exhibit behaviors related to Direct Instruction using Hunter's (1984) Seven Steps of the Teaching Act will lead to global changes in teachers' instructional and questioning behaviors, which in turn will improve student learning in various subjects.	Use of instructional and questioning strategies associated with Direct Instruction and Hunter's (1984) Seven Steps of the Teaching Act (for example, anticipatory set, objective and purpose, instructional input, modeling, checking for guidance)	Author with district support Summer sessions and seven follow-up meetings	About five hours over two months	Group 1: Focused on teaching behaviors applying generically to all subjects
Tienken, 2003 (randomized controlled trial with group equivalence problems)	n/a	There is a need for focused and sustained professional development in writing instruction. Job-embedded professional development or the "environmental" model of instruction will be effective in training teachers in using rubrics to enhance student self-monitoring and thinking about writing and the writing process.	How to provide instruction to students in the use of the criteria in the New Jersey Registered Holistic Scoring Rubric and a set of high-order reflective questions as self-assessment and reflection devices when composing, revising, and editing narrative essays	Author Eight one-hour sessions with six follow-up conferences	14 hours over three and a half months	Group 3: Focused on curriculum and pedagogy, justified by how students learn

n/a is not applicable.

a. "Kennedy content group" refers to the classification in box 3.

Source: Authors' synthesis of studies described in the text.

As professional development research matures, individual empirical studies of multiple professional development programs will eventually make it possible to judge the effectiveness of individual programs

also requires caution.[3] This effect is only a preliminary marker on the sparsely populated terrain of professional development research, still at its developmental stage (Borko, 2004).

Highlighting the problems of many studies of professional development, this report can help researchers avoid methodological pitfalls. Especially important is that researchers undertaking studies with quasi-experimental designs provide data on the baseline equivalence of the treatment and comparison groups. Future studies of the effect of professional development on both teachers and students would be particularly useful—more fully addressing professional development's direct effect on teachers and indirect effect on students.

This report is a first step. As professional development research matures, individual empirical studies of multiple professional development programs will eventually make it possible to judge the effectiveness of individual programs, taking into account such factors as the quality of the study design, statistical significance of the findings, and direction and magnitude of the findings—as does the What Works Clearinghouse classification. Two large-scale impact studies of professional development funded by the Institute of Education Sciences are prime examples of studies under way that can address some of the questions that could not be answered here.

NOTES

1. McCutchen et al. (2002) had 44 teachers and 779 students, but an effect size could be computed only for the kindergarten sample (492 students; the number of kindergarten teachers was not specified).

2. Traditional meta-analysis would weight the studies to account for differences in numbers of effects in each study and the variability in sample sizes across studies. The argument for doing so is that differential weighting affords greater power and precision. The What Works Clearinghouse, however, has not adopted a traditional meta-analysis approach.

3. Following What Works Clearinghouse procedures, this report does not conduct a test of statistical significance on the average effect size, as would have been done in a traditional meta-analysis.

APPENDIX A METHODOLOGY

Developing a review protocol was the first step in systematically reviewing the research-based evidence on the effectiveness of professional development. The approach was modeled on the review process and rigorous evidence standards of the U.S. Department of Education's What Works Clearinghouse. The protocol established the relevance criteria for literature searches and the parameters for screening and reviewing studies (see appendix B for the full protocol and appendix C for key terms and definitions for professional development under the No Child Left Behind Act of 2001). Criteria included:

- *Topic.* The study had to deal with the effects of in-service teacher professional development on student achievement.
- *Population.* The sample had to include teachers of English, mathematics, and science and their students in grades K–12.
- *Study design.* The review of evidence was limited to final manuscripts that were based on empirical studies using randomized controlled trials or a quasi-experimental designs.
- *Outcome.* The study had to measure student achievement outcomes.
- *Outcome measure validity.* The study had to use measures demonstrated to be valid and reliable.
- *Time.* The study had to be published between 1986 and 2006.
- *Country.* Studies had to take place in Australia, Canada, the United Kingdom, or the United States—due to concerns about the external validity of the findings.

A detailed coding guide and a reconciliation form were then developed based on this protocol. The Microsoft Excel–based coding and reconciliation forms were heavily annotated to provide step-by-step, detailed instructions on how to determine and code the relevance, eligibility, and quality of each study. Excel's features and predefined formulas were used in the coding and reconciliation guides to incorporate decision rules stipulated in the protocol. For example, if a study was judged to be a randomized controlled trial and met relevant What Works Clearinghouse evidence standards (lack of problems with randomization, serious attrition, or disruption), the coding guide automatically determined and displayed the quality rating of the study as "met What Works Clearinghouse evidence standards." Excel was programmed to automatically compare the values of coders' entries so that any disagreements would be flagged for review during reconciliation.

Literature searches

Studies were gathered through an extensive electronic search of published and unpublished research literature.[1] The review protocol included a list of keywords that guided the literature search. Seven electronic databases were core data sources: ERIC, PsycINFO, ProQuest, EBSCO's Professional Development Collection, Dissertation Abstracts, Sociological Collection, and Campbell Collaboration. These databases were searched separately for each of the three subjects under review (mathematics, science, and reading and English/language arts). In consultation with a reference librarian, search parameters were developed using database-specific keywords (see appendix D for the list of keywords). A deliberately wide net captured literature on professional development and student achievement, broadly defined. The keyword searches yielded 1,334 studies.

Fourteen key researchers were also asked to identify research for the study. Eight researchers responded, recommending additional studies that fit the study purpose. Finally, existing literature reviews and research syntheses were consulted to ensure that no key studies were omitted. The follow-up literature searches located 25 additional studies, bringing the total to 1,359.

Excluding 16 duplicate records,[2] 1,343 studies were submitted to prescreening. Of these, there were 907 unique studies. The remaining 436 studies were deemed duplicates because they addressed multiple content areas, such as math and science. Because the study was interested in within-content-area findings, that duplication was allowed, and such studies were counted multiple times in the search results (table A1).

Development of the evidence review tool

A Microsoft Access database was designed to facilitate, integrate, and manage the review. The evidence review tool helped centralize and automate such data management and processing tasks as compiling studies from the electronic searches, identifying duplicate records, and collecting and entering full-text documents. The evidence review tool also supported administrative functions such as creating new coding guides and assigning studies to coders for review. Access' built-in functions (queries and report generation) were used to monitor the progress of the review (by study or by coder, for example) and to obtain statistics on the content of the database (the number of studies still missing full-text versions, for example). The evidence review tool made management of the study more efficient and provided easy, hyperlinked access to the full text, coding guide, and reconciliation form for each study.

Coder training. All-day training was conducted for coders in the use of the protocol, coding guide, and reconciliation form. A trainer experienced in the What Works Clearinghouse review process and evidence standards provided the intensive training, using publicly available information from the What Works Clearinghouse website. Coders were also trained in the use of the evidence review tool. Coders met weekly to discuss and resolve issues relevant to the evidence review standards and the rating of the quality of studies.

Screening and coding studies. Six doctorate-level analysts spent four months screening and coding the studies. The screening and coding was conducted in four stages: prescreening, stage 1–full screening, stage 2–coding, and stage 3–coding (figure A1). Appendix E lists the studies that underwent coding in stages 1, 2, and 3.

Prescreening. Because of the wide net, it was expected that keyword searches would yield documents that were not relevant to the report. The prescreening step involved quick scans of abstracts to see if the manuscript met broad relevance and methodology criteria. Coders reviewed manuscripts on five dimensions: focus on K–12 students, focus on at least one of three content areas (math, science, and reading and English/language arts), focus on the effects of teacher professional development, measures of student outcomes, and

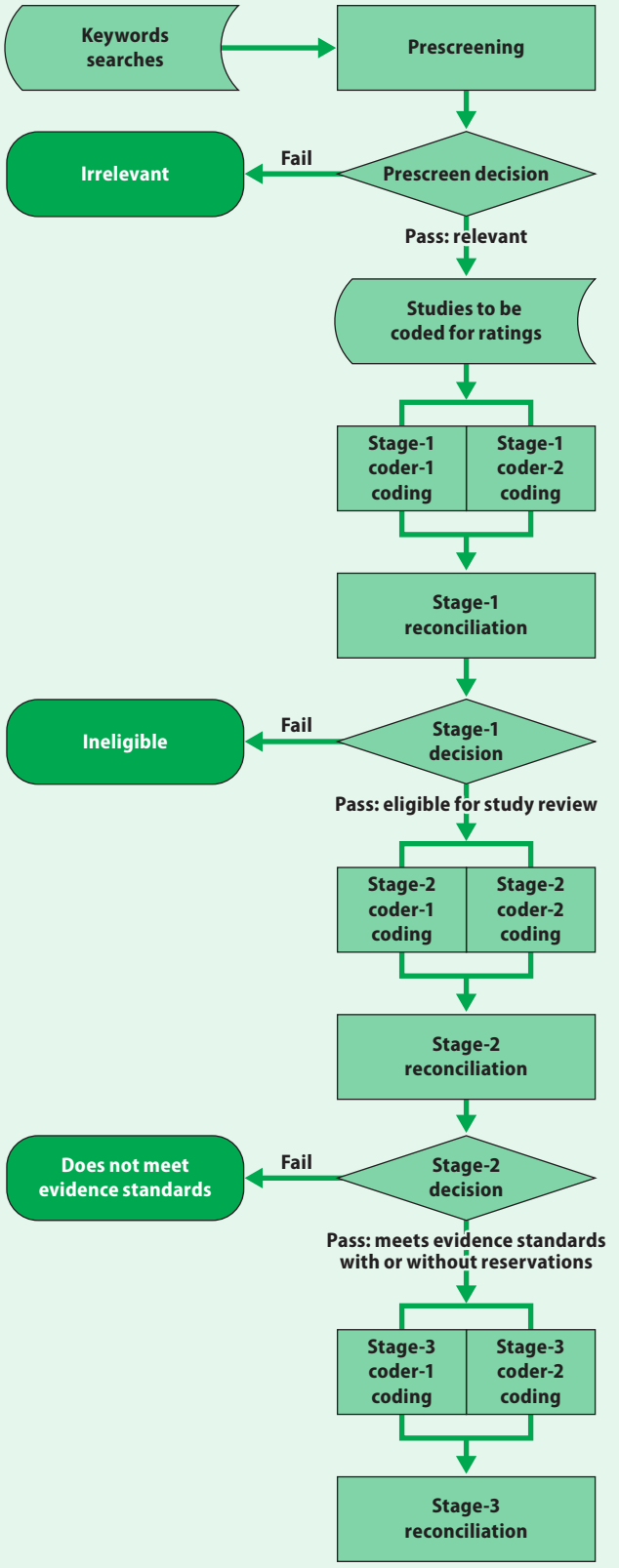
TABLE A1
Number of potentially relevant studies, by subject and data source

Subject	Campbell	Dissertation Abstracts	ERIC	Other[a]	Professional Development Collection (EBSCO)	Proquest	PsycINFO	SocIndex	Subject subtotal
Reading and English/ language arts	31	51	223	5	27	67	52	31	487
Math	27	29	215	10	12	24	24	4	345
Science	48	21	316	10	31	32	40	13	511
Database subtotal	106	101	754	25	70	123	116	48	1,343

a. Sources other than the seven core electronic databases. These were drawn from suggestions by key researchers and literature reviews.

Source: Authors' calculations based on data described in text.

FIGURE A1
Overview of the coding process



Source: Authors' representation of procedures described in text.

empirical and quantitative study design. In cases where the abstract did not provide sufficient information to determine the study's initial relevance, coders sought the full-text version for additional information. Studies that did not meet one or more of these criteria were categorized as "irrelevant" and were excluded from the review.

Of 1,343 studies, 812 were ineligible (slightly more than 60 percent). In many cases, the studies did not focus on professional development. Others were not empirical research but were theoretical papers, opinion pieces, commentaries, conference proceedings, qualitative studies, case studies, literature reviews, research syntheses, or meta-analyses.

The next most frequent reason for failing the prescreening was the lack of a student achievement outcome measure (800 studies). Lack of focus on the effects of teacher professional development was the third most common reason. It appears that keyword searches successfully filtered studies for K–12 grade relevance and target-subject relevance. Fewer studies missed on these two criteria (table A2).

Only 132 unique studies met all five criteria in the prescreening step and were sent to the next stage of review process.

Stage 1–full screening. Stage 1–full screening was a more detailed version of the prescreening. Pairs

TABLE A2
Number and share of studies failing to meet the prescreening criteria

Prescreening criterion	Number	Percentage
Focus on K–12 students	349	25.9
Focus on target subjects	518	38.6
Focus on the effects of teacher professional development	761	56.7
Measuring student achievement outcomes	800	59.6
Quantitative and empirical study	812	60.5

Source: Authors' calculations based on data described in text.

of coders independently read full-text versions of the studies and rated each study on eight criteria. Inter-rater reliability for the fail reasons in stage 1 was excellent, ranging from 83 percent to 98 percent, with the overall agreement rate at 92 percent. At the end of the double-coding, coders held a reconciliation session to resolve any disagreements.

There were eight relevance criteria in stage 1: study topic (in-service professional development), sample (K–12 teachers and their students), country (Australia, Canada, United Kingdom, or United States), time of study (1986 or later), study design (randomized controlled trials or quasi-experimental designs), student achievement outcome measure in the specified subjects, focus on the effects of in-service professional development on student achievement, and psychometric properties of student outcome measures. Studies that did not meet one or more of these criteria were categorized as "ineligible for study ratings review" and did not pass stage 1 screening. Twenty-seven studies (20 percent) met all the stage 1 criteria and were eligible for continuation to stages 2 and 3 (table A3).

Most of the studies (84 of 132, or about 64 percent) failed to meet the rigorous study design criterion. Only 48 studies met this criterion. Half were randomized controlled trials, and half quasi-experimental designs. The lack of focus on the effects of in-service professional development on student achievement was the next most common reason that studies were excluded (for 38 studies, a distant second at just under 29 percent).

The rate of agreement between the trained coders ranged from lows of 83 percent (focus on of in-service professional development) and 84 percent (study design relevance) to a high of 98 percent (K–12 grade and country relevance).

Stage 2 coding. The 27 studies that passed the stage 1 coding went to stage 2. As in stage 1, pairs of coders read and rated each study independently, then met with a third coder, who resolved all disagreements. Inter-rater reliability for the individual fail reasons was good, ranging from 65 percent (judging the baseline equivalence of quasi-experimental designs) to 100 percent (judging whether a randomized controlled trial had a randomization problem), with the overall agreement rate at 77 percent. This lower reliability was expected because of the greater technicality of this stage of review. Disagreements were resolved during the reconciliation session.

TABLE A3
Studies failing and passing stage 1 criteria

	Failing		Passing	
Stage 1–full screening criterion	Number	Percentage	Number	Percentage
Focus on in-service professional development	30	22.7	102	77.3
Focus on K–12 teachers and their students	2	1.5	130	98.5
Country	8	6.1	124	93.9
Time of study	13	9.9	119	90.1
Study design	84	63.6	48	36.4
Focus on the specified subjects	24	18.2	108	81.8
Focus on the effects of in-service professional development on student achievement outcomes	38	28.8	94	71.2
Overall stage 1 screening decision	105	79.5	27	20.5

Note: Each row contains 132 studies. Questions about adequate psychometric properties were asked only if all seven preceding criteria were met. Because not all 132 studies were subject to that question, it is excluded from this table.

Source: Authors' calculations based on data described in text.

At this stage, coders determined the evidence of causal validity in each study according to What Works Clearinghouse evidence standards and gave each study one of three ratings: "meets evidence standards" (for randomized controlled trials that provided the strongest evidence of causal validity), "meets evidence standards with reservations" (for quasi-experimental studies and randomized controlled trials that had problems with randomization, attrition, or disruption), and "does not meet evidence screens" (for studies that did not provide strong evidence of causal validity).

Of the 27 studies, 7 were randomized controlled trials and 20 were quasi-experimental designs. Only nine studies met evidence standards and were submitted to the final stage of coding.

Of the 18 studies that did not meet evidence screens, 17 were quasi-experimental designs and 1 was a randomized controlled trial. Sixteen of the quasi-experimental designs had problems with baseline equivalence between groups. In many cases, these studies failed to collect any baseline measures, such as pretest outcome scores. In others, initial baseline differences between intervention and comparison groups were too large to be accounted for by any statistical method. One quasi-experimental design was excluded because of high attrition. The excluded randomized controlled trial had problems with both attrition and baseline equivalence.

Stage 3 coding. The nine studies that met evidence standards or met evidence standards with reservations were reviewed further to describe other characteristics of the study and of the professional development (tables A4 and A5). These characteristics included:

- Estimated impact of the professional development (in effect sizes and improvement indices).
- Replicability of the professional development and the study.
- Teacher outcome measures.
- Content and form of the professional development (using the classification in Kennedy, 1998) and other professional development–related features, such as duration and intensity.
- Whether the effect of professional development was confounded with that of curriculum.
- Statistical analysis.
- Statistical reporting.

Notes

1. Unlike What Works Clearinghouse reviews, this report did not seek submissions from intervention developers and the public.
2. These were duplicates within a single subject domain, typically uncovered by two or more databases.

TABLE A4

Basic features of the nine studies that meet evidence standards

Study	Study design	Content area	School level	Student outcomes examined
Carpenter et al., 1989	Randomized controlled trial	Mathematics	Elementary (1st grade)	Students' computation and math problem-solving scores on the Iowa Test of Basic Skills, Level 7
Cole, 1992	Randomized controlled trial	Mathematics and reading and English/ language arts	Elementary (4th grade)	Students' mathematics, reading, and language test scores on the Stanford Achievement Test
Duffy et al., 1986	Randomized controlled trial	Reading and English/ language arts	Elementary (5th grade)	Students' reading comprehension test scores on the Gates-MacGinitie tests
Marek & Methven, 1991	Quasi-experimental design	Science	Elementary (K–3rd, 5th grades)	Students' conservation reasoning, as measured by Piagetian cognitive tasks
McCutchen et al., 2002	Quasi-experimental design	Reading and English/ language arts	Elementary (K–1st grades)	Students' alphabetics (Test of Phonological Awareness), orthographic fluency (a timed alphabetic writing task), comprehension (the comprehension subtest of the Metropolitan Readiness Tests), word reading (Gates-MacGinitie Reading Tests), and writing skills (a composition task)
McGill-Franzen et al., 1999	Randomized controlled trial	Reading and English/ language arts	Elementary (kindergarten)	Students' receptive language skills (the Peabody Picture Vocabulary Test) and early literacy skills (subtests of the Concepts about Print and Diagnostic Survey)
Saxe et al., 2001	Quasi-experimental design	Mathematics	Elementary (4th–5th grades)	Students concepts and computation of fractions, as assessed by a 29-item, 40-minute timed measure developed by the authors
Sloan, 1993	Randomized controlled trial	Mathematics, science, and reading and English/ language arts	Elementary (4th–5th grades)	Students' reading, math, and science scores, measured by the Comprehensive Test of Basic Skills
Tienken, 2003	Randomized controlled trial with group equivalence problems	Reading and English/ language arts	Elementary (4th grade)	Students' narrative writing, as measured by content/organization scores on a standardized writing test administered as part of New Jersey's Elementary School Proficiency Assessment

Source: Authors' synthesis of studies described in text.

TABLE A5

Brief descriptions of the nine studies that meet evidence standards

Study (study design)	Description
Carpenter et al., 1989 (randomized controlled trial)	Forty first-grade teachers were randomly assigned to participate in a month-long workshop on children's development of problem-solving skills in addition and subtraction (n = 20; see table 3 for additional details). The control group teachers participated in two two-hour workshops during the instructional year. These workshops were intended to provide control teachers reinforcement for their participation in the study, not to create a contrasting treatment group. Unlike in the intervention group's workshop, no mention was made of how children think as they solve problems. Instead, the focus was on the use of nonroutine problems to motivate students to engage in problem-solving. Data collected at the teacher level included classroom observations and measures of teacher knowledge and beliefs. Twelve students (six girls and six boys) were randomly selected from each class to provide data on student outcomes. Students with special learning needs were omitted from the random selection. Data collected at the student level included a standardized mathematics achievement test (Iowa Test of Basic Skills, ITBS) and an interview to assess students' problem-solving strategies. The researchers also administered three math achievement scales constructed from combinations of items from ITBS items and researcher-developed items. Because of the overlap in the ITBS scores and the three researcher-constructed scales, only the ITBS scores are reported in this report. The student problem-solving strategies interview is also omitted from the analyses here because there was no direct measure of student achievement. The authors found no statistically significant difference between the treatment and control groups on the student outcome measures, but both were positive (favoring the treatment group) and large enough to be considered substantively important. This study was judged to be a randomized controlled trial that met What Works Clearinghouse standards.
Cole, 1992 (randomized controlled trial)	Twelve fourth-grade teachers and their intact classes in an intermediate school in Mississippi were randomly assigned into treatment and control groups. The six treatment teachers underwent a comprehensive staff development training program using Mississippi Teacher Assessment Instrument modules for training materials (see table 3). No details were provided about the control group teachers or any professional development they may have had. No teacher outcome measures were gathered, but classroom observations were done in the six treatment classrooms to assess fidelity of implementation. Students' math, reading, and language scores on the Stanford Achievement Test were the outcome measures (for 268 students). Students' third-grade test scores from the spring of 1989 were used as pretests, and their fourth-grade test scores from the spring of 1990 were used as the post-tests. Results were reported by eight student subgroups (combinations of low and high socioeconomic status, black and white, and male and female), and the author reported statistically significant differences on 10 comparisons of the 24. This report applies corrections to the statistical significance of the results reported by the author to adjust for unaddressed clustering and for multiple outcomes. For comparability with the other studies, the average effect size and improvement index are reported for each content domain (math, reading, and language), summed across all eight student subgroups. The average effects in math and the reading were positive (favoring the treatment group) and statistically significant, according to the analysis for this report. The average effect in language was positive but not large enough to be considered substantively important. This study was judged to be a randomized controlled trial that met What Works Clearinghouse standards.

(CONTINUED)

TABLE A5 (CONTINUED)

Brief descriptions of the nine studies that meet evidence standards

Study (study design)	Description
Duffy et al., 1986 (randomized controlled trial)	Twenty-two fifth-grade teachers and their intact classes were randomly assigned into equal-sized treatment and control groups. The professional development received by the treatment group teachers focused on explicit instructional talk (see table 3). Control group teachers attended a presentation on effective classroom management. The teachers were unaware that the two groups received different training. Classroom observations were conducted four times during the school year to document instructional practices in the two types of classrooms. The study took place in a large urban district that implemented a policy of using the Joplin Plan to group students homogenously for reading. Within each classroom, students were identified as low-achieving readers based on their fourth-grade Stanford Achievement Test scores and fourth-grade teachers' recommendations. All the low-achieving readers scored more than one year below grade level in reading. The number of students in the low-achieving reader groups ranged from 4 to 22, with an average group size of 11.8 (259 students were included, 130 in the treatment group and 129 in the control). This study's student-level outcomes focused only on the achievement of students in the low-achieving groups in the 22 classrooms, as measured by pretest and post-test administrations of the Gates-MacGinitie Reading Test. Also administered was a student strategy awareness measure, not included in the results in this report because it was not an achievement outcome. The authors found no statistically significant differences in students' Gates-MacGinitie scores. This study was judged to be a randomized controlled trial that met What Works Clearinghouse standards.
Marek & Methven, 1991 (quasi-experimental design)	Sixteen elementary school teachers applied for and participated in a National Science Foundation–sponsored workshop that focused on science as knowledge and knowledge-seeking and how to develop a curriculum of learning cycles that represented this philosophy (see table 3). Eleven comparison group teachers were identified through a nomination procedure, with the intervention group participants asked to identify teachers in their schools who were the same gender, taught the same grade, had similar teaching experience, and who taught science by exposition. Teachers taught kindergarten, first grade, second grade, third grade, and fifth grade. Classroom observations were conducted to document instructional practices in the two types of classrooms. Ten students from each of the 27 teachers' classrooms were randomly selected and interviewed to assess conservation reasoning. Three Piagetian conservation tasks (liquid amount, weight, and length) were given at the beginning and the end of the school year. If a student was able to conserve on a task, a score of one was recorded. So, each child could score from zero to three. No significant differences between groups was found on pretest conservation, but the authors reported statistically significant differences on total conservation post-test scores for the third graders. This report applies a correction to the statistical significance of the result reported by the author to adjust for unaddressed clustering, finding a positive and statistically significant effect favoring the treatment group. This was judged to be a quasi-experimental design study that met What Works Clearinghouse standards with reservations.

Study (study design)	Description
McCutchen et al., 2002 (quasi-experimental design)	Forty-four kindergarten and first-grade teachers responded to an invitation to participate in the study. A total of 43 classrooms (23 treatment and 20 comparison) were followed, because two of the treatment-group teachers teamed in the same classroom. The professional development given to the treatment-group teachers focused on deepening teachers' knowledge of phonology and its link to orthography (see table 3). Several survey measures of teacher knowledge were administered, and classroom observations were done in all the classrooms to record teachers' literacy instruction. A total of 779 students responded to multiple measures of early reading and writing skills (see table A4). The analysis sample consisted of 492 kindergarteners (268 in the treatment group and 224 in the comparison group) and 287 first graders (157 in the treatment group and 130 in the comparison group). Although multiple measures of students' achievement were administered, the authors did not report enough detail about their analyses to allow this report to compute effect sizes for the entire sample. So, an effect size is calculated only for the Gates-MacGinitie word reading subtest of the kindergarten sample. To avoid discarding the study, that result is included here. The authors reported positive, statistically significant results favoring the treatment group. No clustering adjustment to the statistical significance of the finding was necessary because of the hierarchical analyses. This was judged to be a quasi-experimental design study that met What Works Clearinghouse standards with reservations.
McGill-Franzen et al., 1999 (randomized controlled trial)	Eighteen kindergarten teachers, three each from six schools, were randomly assigned into one of three groups: training and books (the treatment group), no training and books, and no training and no books. This report presents results comparing training-and-books teachers with no-training-and-no-books teachers. The professional development consisted of techniques for encouraging children to pick up books and read them (see table 3). The authors collected three types of data to measure classroom environment: classroom observations, teacher interviews, and teacher weekly read-aloud logs. The primary outcomes of this study were at the student level (with 317 students, 164 treatment and 153 control). Children's early literacy and writing skills were measured using a variety of standardized tests (see table A4), administered at the beginning and the end of the school year. The authors reported positive, statistically significant differences on all measures except the Peabody Picture Vocabulary Test. This report applies corrections to the statistical significance of the other five results reported by the authors to adjust for unaddressed clustering and for multiple outcomes. Three of the results remain positive and statistically significant (concepts about print, letter identification, and hearing sounds in words), and two effects are substantively important but not statistically significant (writing vocabulary and Ohio Word Test). This study was judged to be a randomized controlled trial that met What Works Clearinghouse standards.
Saxe et al., 2001 (quasi-experimental design)	Twenty-three teachers in the Los Angeles area responded to an invitation to participate in this year-long study. Based on teachers' responses to a prescreening questionnaire, three groups were formed. The Integrated Mathematics Assessment (IMA), was the treatment condition (with nine teachers, and the Collegial Support (SUPP, eight teachers), and Traditional Instruction (TRAD, six teachers) groups were the comparison groups. This report presents results comparing the IMA and TRAD groups. The professional development focused on enhancing teachers' understanding of fractions, student cognition, and student motivation (see table 3). The authors did not collect any teacher-level data. The student outcome measures were two researcher-developed tests of fraction concepts and of fraction computations, administered at the beginning and the end of the school year. The authors conducted analyses of covariance on the classroom-level data and found no statistically significant differences between the IMA and TRAD groups on the computational scale, but the effect was negative (favoring the TRAD group) and large enough to be considered substantively important. The authors found strong and statistically significant differences between the groups on the fraction concepts measure, favoring the IMA group. This was judged to be a quasi-experimental design study that met What Works Clearinghouse standards with reservations.

(CONTINUED)

TABLE A5 (CONTINUED)

Brief descriptions of the nine studies that meet evidence standards

Study (study design)	Description
Sloan, 1993 (randomized controlled trial)	Ten fourth- and fifth-grade teachers in seven Midwestern schools were randomly assigned to two conditions: Direct Instruction training and a control group. Teachers in the treatment group were trained to use the questioning and instructional behaviors associated with the Direct Instruction model (see table 3). No details were provided about the control group teachers and any professional development they may have had. Classroom observations were conducted to document the instructional environments in both types of classrooms. The seven fourth-grade and the three fifth-grade classrooms contained 173 students. The Comprehensive Test of Basic Skills was administered as pretest and post-test, measuring students' achievement in reading, mathematics, science, and social studies. Self-esteem and classroom environment were also measured, but they are not included in this report because they are not achievement outcomes. The social studies outcomes are also excluded because social studies was not among the content areas in the protocol. The author found no statistically significant differences between groups on the Comprehensive Test of Basic Skills mathematics score but reported statistically significant results favoring the Direct Instruction group on the reading and science scores. This report applies corrections to the statistical significance of these two results to adjust for unaddressed clustering and for multiple outcomes and finds that neither effect is statistically significant. But both are still large enough to be considered substantively important. This study was judged to be a randomized controlled trial that met What Works Clearinghouse standards.
Tienken, 2003 (randomized controlled trial with group equivalence problems)	This small, post-test-only randomized trial involved five fourth-grade teachers and their 98 students in a New Jersey school. Two teachers were trained to teach students to use scoring rubrics and reflective questions as self-assessment devices (see table 3). No details were provided about the control group teachers and any professional development they may have had. Treatment group teachers were asked to complete reflective logs and their classrooms were observed as measures of implementation fidelity. At the end of the school year students' content/organization scores on the state's standardized writing assessment were compared. The author reported a positive, statistically significant difference favoring the treatment group. This report applies a clustering correction and finds that the result is no longer statistically significant. However, the effect is large enough to be considered substantively important. Because of the post-test-only design, the teacher randomization was insufficient to ensure that students in the five classrooms were comparable in their baseline writing skills. Therefore, this study was judged to be a randomized controlled trial with group equivalence problems that met What Works Clearinghouse standards with reservations.

Source: Authors' synthesis of studies described in text.

APPENDIX B
PROTOCOL FOR THE REVIEW OF RESEARCH-BASED EVIDENCE ON THE EFFECTS OF PROFESSIONAL DEVELOPMENT ON STUDENT ACHIEVEMENT

Developed for Regional Education Laboratory–Southwest by American Institutes for Research

IES Approved, December 6, 2006

Abstract

Topic area focus. As part of the Southwestern Regional Educational Laboratory's (REL Southwest) fast-turnaround projects, the American Institutes for Research (AIR) will conduct a systematic review of research-based evidence on the effects of professional development on growth in student learning. The main focus of the review will be how students' achievement in three core academic subjects (English/language arts/reading, mathematics, and science) is affected by professional development activities that are designed to enhance K–12 teachers' knowledge and skills and to transform their classroom practices.

A basic assumption of this review is that the effects of professional development on student achievement are mediated by increased teacher knowledge and improved teaching in the classroom (see appendix B, figure B.1). Existing literature reviews (Loucks-Horsley & Matsumoto, 1999; Supovitz, 2001) indicate that the volume of literature on the effect of professional development on student learning is thinner than that on the effects of professional development on teacher learning and classroom teaching practices. Therefore, we expect that our literature search will turn up existing studies on the effects of professional development on teacher learning and teaching practice (but which fall short of demonstrating its effect on student achievement), as well as those that take the next step and address the link between professional development and student outcomes. Our tally of excluded studies will be the means by which we document the paucity of research that directly examines the effect of professional development on student achievement.

This systematic review of evidence will address the following research questions:

- What is the impact of providing professional development to teachers on student achievement? If a sufficient number of studies remain in the final pool, we will also try to disaggregate the results to answer:
  - Does the effect of teacher professional development on student achievement vary by type of professional development provided (for example, summer institutes, workshops, online training)?
  - Does the effect of teacher professional development on student achievement vary by content domain (English/language arts, mathematics, science)?
  - Does the effect of teacher professional development on student achievement vary by grade level (elementary, secondary)?

General inclusion criteria

Populations to be included. Target populations for this review include the students of K–12 teachers of English/language arts/reading, mathematics, and science. Although we would like to be able to examine how the effect of teacher professional development on student achievement varies by student characteristics (for example, English language learners, economically disadvantaged students, students with disabilities), we do not expect to find many studies that directly address student outcomes, which are distal effects of professional development given to teachers. If our final review pool contains studies that allow for this disaggregation, we will include those findings in the final report.

Types of professional development to be included. The No Child Left Behind provisions shed light on

what constitutes professional development (see appendix C for detailed definitions). It encompasses a wide range of activities that are designed to provide teachers with opportunities to deepen their knowledge in the subject matter that they teach, improve teaching skills, and better understand how students learn and think.

Therefore, we take an inclusive view on the form and substance of professional development (Kennedy, 1998). A variety of forms (format and structure) and substances (content and purpose) of professional development will be considered for the inclusion of review as long as they are designed to assist teachers of English/language arts/reading, mathematics, and science to achieve their desired goals for enhancing student achievement outcomes.

- The substance of professional development may include combinations of the following areas:
    - Research-based reform models, curricula, instructional strategies and models, or materials (for example, Cognitively Guided Instruction, America's Choice, Open Court, Success for All)
    - Content knowledge (for example, phonemic awareness, algebraic concepts, use of manipulatives, conservation)
    - Pedagogical content knowledge of a particular subject: knowledge about how students learn a particular subject and understanding of student thinking
    - Generic instructional strategies or teaching skills that are applicable to any subject (for example, differentiated instruction, cooperative learning, and reciprocal learning); this may include such special topics as classroom management, use of assessment data, alignment of instruction with standards, and teaching students with special needs in learning English, mathematics, or science (for example, English language learners and students with disabilities).
- The form of professional development to be included in the review may involve:
    - Traditional types of professional development such as workshops, summer institutes, and conferences.
    - Reform types of professional development, such as coaching and mentoring, that are embedded in teachers' classroom teaching.
    - Online professional development such as online courses, web-based teaching modules, or virtual teacher-learning communities.

Types of research studies to be included. Our review of professional development literature focuses on studies that involve student learning in reading, mathematics, and science in grades K–12. To be included in the review, a study must meet several relevancy criteria:

- *Topic.* The study has to deal with professional development applied to teaching in reading, mathematics, and science. The study is required to focus on the effects of teachers' in-service professional development on student learning. Hence, this review does not include studies that are primarily focused on:
    - Effects of pre-service teacher preparation on student learning.
    - Effects of teacher quality in general on student achievement.
    - Effects of comprehensive reform models, curricula, instructional models, materials, and assessment on student achievement, with little attention to professional development (for example, teacher

training being provided as part of technical assistance).

- Properties of measurement instruments (for example, developing measures of teacher's content knowledge).

- Policy analysis (for example, studies that describe the implementation and impact of such reform policies as the National Science Foundation's Systemic Initiatives or Math-Science Partnership program).

- *Time.* The review of the evidence on professional development and student achievement focuses on a 20-year span, from 1986 to 2006. However, we may include the following studies on a case-by-case basis:

  - Seminal studies identified by key researchers in the field, regardless of the year of publication.

  - Some work in progress involving a multiyear longitudinal study design (for example Institute of Education Sciences–funded professional development impact studies) merits special attention. These ongoing studies may not be included in our review during the current study period (for example, interim reports; note that we will not accept any manuscript labeled as "draft"). However, given the significance of these studies, it is important to review in a timely manner any emerging evidence from the studies. Hence, we offer the option to update our review on a yearly basis to include any newly published reports from the recent multiple-year studies, provided that an extension in contract is granted with supplemental funds.

- *Sample.* The sample must include teachers of English, mathematics, and science and their students in grades K–12.

  - Pre-service teachers are not included in this review. In addition, teachers of other academic subjects are also not included.

- *Study design.* The study design and focus are limited to final manuscripts that:

  - are empirical studies, using quantitative methods and inferential statistical analysis, and

  - take the form of a randomized controlled trial or a quasi-experimental design.

- *Outcome.* The study is required to focus on student outcomes of professional development.

  - Student outcomes must involve academic achievement in reading, mathematics, or science (e.g., reading score gains in state assessments). Even though other student outcomes such as positive attitude toward the subject they learn, motivation, and self-efficacy are important outcomes on their own right, they are not the focus of our review.

    Student outcomes in reading, math, or science may include the following:

    - English/language arts/reading: Phonemic awareness, phonological awareness, print awareness, letter knowledge, phonics, reading fluency, vocabulary development, reading comprehension, grammar, writing, communication, and critical thinking.

    - Mathematics: Number sense, operations, geometric concepts, algebraic concepts, measurement, data analysis; skills in performing procedures, logical reasoning, and solving nonroutine problems.

    - Science: knowledge in earth science, life science, and physical science,

science inquiry skills, scientific reasoning, science experiment design, data interpretation and analysis, hypothesis testing, and explanation formulation from evidence.

Specific study parameters

The following parameters specify which studies are to be considered for review and which aspects of those studies are to be coded for the review:

1. *Validity and reliability of outcome measures.* Study must include at least one relevant outcome measure that meets minimum requirements for face validity or reliability. For example, if a study presents a measure that does not have face validity or has some measure of reliability (for example, Cronbach's alpha), the measure would be excluded; if that measure was of the only relevant outcome, the entire study would be excluded.

2. *Characteristics relevant to equating groups.* Important contextual factors as well as pre-existing teacher quality and student characteristics that might be related to the outcomes of professional development must be equated if a study does not employ random assignment as part of its design. Such pre-existing factors include:

    - School and classroom contexts under which in-service professional development is undertaken (for example, small learning community, teacher learning community, trust in schools).
    - Pretest measures of teachers' beliefs, knowledge, skills, or instructional practices.
    - Individual characteristics and qualifications of teachers, such as teaching experience, degree, and major.
    - Pretest measures of students' achievement in reading, mathematics or science.
    - Individual or demographic characteristics of students such as intelligence quotient, socioeconomic status, and special learning needs

The issue of when the equating was done must also be considered, as well as whether the equating procedure may have resulted in groups with extreme scores in measurements (because upon repeated measurements, these scores tend to move toward the average, even without an intervention).

3. *Effectiveness of professional development across different groups.* The effect of professional development on student achievement may vary by student characteristics. A study may examine the effects of professional development within important student subgroups, which may include:

    - Students with different learning styles, students with disabilities, students with special learning needs (including students who are gifted and talented), and students with limited English proficiency.
    - Students of differing achievement levels (for example, poor readers, underachievers)
    - Students who are ethnic or racial minorities.

4. *Effectiveness of the professional development across different settings and contexts.* The effectiveness of professional development on student achievement may also vary by settings. A study may examine the effects of professional development across different settings. These settings may include:

    - School or class size.
    - School-level poverty and minority concentration level.
    - School location (urban, rural, suburban).

- School improvement status under No Child Left Behind.
- Classroom types (for example, general education or special education, inclusion classrooms)

5. *Measuring post-intervention effects.* There exists a window of opportunity to observe the effects of professional development. A time lag between the enactment of professional development (as intervention) and the measurement of its effects on teacher and student learning may range from days to weeks to months, or even to years. The optimal time lapse between the implementation of professional development and the measurements of outcomes may vary by the nature of professional development as well as by the nature of the outcomes. For example, if the implementation of professional development requires teachers' sustained participation followed by ongoing supports (for example, peer coaching as opposed to short-term workshops), it requires an extended time lapse between the beginning of the intervention and the post-intervention outcome measurement. Further, determining the effectiveness of professional development would require a longer time interval for student learning (as a distal outcome) than for teacher learning (as a proximal outcome). At any rate, it is important to document when post-intervention effects were measured to determine whether a sufficient time lapse was provided to observe any significant effect of professional development.

6. *Defining attrition.* The burden is on the study authors to demonstrate post-attrition group equivalence on pretest measures both for overall attrition and for differential attrition between study groups. Post-attrition group equivalence must be shown through either a well-powered (0.80) test of equivalence that is nonsignificant or a standardized mean difference between groups of less than $d = 0.10$.

7. *Avoiding confounding teacher and intervention effects.* In a randomized controlled trial or a quasi-experimental design study, there should be more than one teacher assigned to each condition. A teacher–intervention confound occurs when only one teacher assigned to each condition. If a teacher–intervention confound exists, the study may be excluded or downgraded. The final judgment of the study quality will depend on the details of the study, such as demonstration of negligible teacher effects, methods for teacher or student assignment, or the appropriateness of the equating procedures.

8. *Statistical properties important for computing accurate effect sizes.* For most statistics (including *d*-indices), normal distribution and homogenous variances are important properties. For odds ratios there are no required desirable properties except the minimum of five observations per cell.

In the cases where effect sizes do not reach statistical significance, we consider an effect size equal to or greater than |0.25| as the minimum threshold for judging an intervention to have had an effect. The value of 0.25 corresponds to a 10 percentile point difference between the mean of the control group (fiftieth percentile) and the mean of the intervention group (sixtieth percentile) on a normal distribution.

In the case where a misaligned analysis is reported (the unit of analysis is not the same as the unit of assignment) and the author is unable to provide a corrected analysis, the effect sizes computed will incorporate a statistical adjustment for clustering. According to the standards determined by the What Works Clearinghouse Technical Advisory Committee, the default intra-class correlation used for achievement outcomes is 0.20.

Methodology

Collecting and screening studies. The literature search is intended to be comprehensive and

systematic. A detailed protocol that includes a list of keywords (see appendix D) guides the entire literature search process. At the beginning of the process, relevant journals, organizations, and experts are identified. AIR will search core sources and additional topic-specific sources identified by the content experts. Next, by using a well-defined coding guide, AIR will screen and code studies that are collected with the literature search.

Sources for studies. Trained AIR staff members will use the following strategies to search electronic databases and the "fugitive" or "gray" literature:

Search of electronic databases. These electronic databases will be searched:

1. *ERIC.* Funded by the U.S. Department of Education, ERIC is a nationwide information network that acquires, catalogs, summarizes, and provides access to education information from all sources. All U.S. Department of Education publications are included in its inventory.

2. *PsycINFO.* PsycINFO contains more than 1.8 million citations and summaries of journal articles, book chapters, books, dissertations, and technical reports, all in psychology. Journal coverage, which dates back to the 1800s, includes international material selected from more than 1,700 periodicals in more than 30 languages. More than 60,000 records are added each year.

3. *Wilson Education Abstracts PlusText.* Wilson Education Abstracts PlusText, also known as Education PlusText, combines abstracts and indexing from H.W. Wilson's Education Abstracts database with thousands of full-text and full-image articles. The database includes indexing and abstracts for articles published by more than 400 journals cited in H.W. Wilson's Education Abstracts database. It also includes full-text and full-image coverage for more than 175 of the sources. Overall dates of coverage are 1994 to the present. Special education, adult education, home schooling, and language and linguistics are just a few of the hundreds of topics users can research in the database.

4. *Professional Development Collection.* Designed for professional educators, this database provides a highly specialized collection of more than 500 full-text journals, including nearly 350 peer-reviewed titles. Professional Development Collection is the most comprehensive collection of full-text education journals in the world.

5. *Dissertation Abstracts.* As described by Dialog, Dissertation Abstracts is a definitive subject, title, and author guide to virtually every American dissertation accepted at an accredited institution since 1861. Selected master's theses have been included since 1962. In addition, since 1988, the database includes citations for dissertations from 50 British universities that have been collected by and filmed at The British Document Supply Center. Beginning with Dissertation Abstracts International, Volume 49, Number 2 (Spring 1988), citations and abstracts from Section C, Worldwide Dissertations (formerly European Dissertations), have been included in the file. Abstracts are included for doctoral records from July 1980 (Dissertation Abstracts International, Volume 41, Number 1) to the present. Abstracts are included for master's theses from Spring 1988 (Masters Abstracts, Volume 26, Number 1) to the present.

6. *Sociological Collection.* This database provides coverage of more than 500 full-text journals, including nearly 500 peer-reviewed titles. Sociological Collection offers information in all areas of sociology, including social behavior, human tendencies, interaction, relationships, community development, culture, and social structure. This database is updated daily via EBSCOhost.

7. *Campbell Collaboration.* C2-SPECTR (Social, Psychological, Educational, and Criminological Trials Register) is a registry of more than

10,000 randomized and possibly randomized trials in education, social work and welfare, and criminal justice.

In consultation with the AIR librarian, search parameters will be developed with the use of database-specific keywords (see appendix D for the preliminary list of keywords).

Search of "fugitive" or "gray" literature. Our search for fugitive or grey literature encompasses the following strategies:

1. Solicitations are made to key researchers ("snowballing" approach).

2. Checking prior literature reviews and research syntheses (using the reference lists of prior reviews and research syntheses to make sure we have not omitted key studies).

Protocol references

Ball, D. L., & Cohen, D. K. (1999). Developing practices, developing practitioners: Toward a practice-based theory of professional development. In G. Sykes & L. Darling-Hammonds (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 30–32). San Francisco, CA: Jossey-Bass.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 30*(8), 3–15.

Cohen, D. & Hill, H. C. (1998). *Instructional policy and classroom performance: The mathematics reform in California*. CPRE Research Report Series RR-39. Philadelphia: Consortium for Policy Research in Education.

Cohen, D. K., Raudenbush, S., & Ball, D. L. (2002). Resources, instruction, and research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research*, (pp. 80–119). Washington, DC: Brookings Institution Press.

Elmore, R. F. (1997). *Investing in teacher learning: Staff development and instructional improvement in Community School District #2, New York*. Philadelphia: Consortium for Policy Research in Education.

Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Education Research Journal, 38*(4), 915–945.

Guskey, T. (2003). What makes professional development effective? *Phi Delta Kappan*, *84*(10), 748–750.

Kennedy, M. (1998). *Form and substance of inservice teacher education*. (Research Monograph No. 13.) Madison, WI: National Institute for Science Education, University of Wisconsin–Madison.

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

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.

Showers, B., Joyce, B., & Bennett, B. (1987). Synthesis of research on staff development: A framework for future study and a state-of the-art analysis. *Educational Leadership*, *45*(3), 77–87.

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

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

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.

APPENDIX C
KEY TERMS AND DEFINITIONS RELATED TO PROFESSIONAL DEVELOPMENT

According to the provisions of the No Child Left Behind Act of 2001 (section 9101 under part A of title IX), the term *professional development*:

(A) Includes activities that

(i) Improve and increase teacher's knowledge of the academic subjects the teachers teach, and enable teachers to become highly qualified;

(ii) Are an integral part of broad schoolwide and districtwide educational improvement plans;

(iii) Give teachers, principals, and administrators the knowledge and skills to provide students with the opportunity to meet challenging state academic content standards and student academic achievement standards;

(iv) Improve classroom management skills;

(I) Are high quality, sustained, intensive and classroom-focused in order to have a positive and lasting impact on classroom instruction and the teacher's performance in the classroom;

(II) Are not one-day or short-term workshops or conferences;

(vi) Support the recruiting, hiring, and training of highly qualified teachers, including teachers who became highly qualified through state and local alternative routes to certification;

(vii) Advance teacher understanding of effective instructional strategies that are:

(I) Based on scientifically based research (except that this subclause shall not apply to activities carried out under part D of title II); and

(II) Strategies for improving student academic achievement or substantially increasing the knowledge and teaching skills of teachers; and

(viii) Are aligned with and directly related to:

(I) State academic content standards, student achievement standards, and assessments; and

(II) The curricula and programs tied to the standards described in subclause (I) except that this subclause shall not apply to activities described in clauses (ii) and (iii) of section 2123(3)(B);

(ix) Are developed with extensive participation of teachers, principals, parents, and administrators of schools to be served under this Act;

(x) Are designed to give teachers of limited English proficient children, and other teachers and instructional staff, the knowledge and skills to provide instruction and appropriate language and academic support services to those children, including the appropriate use of curricula and assessments;

(xi) To the extent appropriate, provide training for teachers and principals in the use of technology so that technology and technology applications are effectively used in the classroom to improve teaching and learning in the curricula and core academic subjects in which the teachers teach;

(xii) As a whole, are regularly evaluated for their impact on increased teacher effectiveness and improved student academic achievement, with the findings of the evaluations used to improve the quality of professional development;

(xiii) Provide instruction in methods of teaching children with special needs;

(xiv) Include instruction in the use of data and assessments to inform and instruct classroom practice; and

(xv) Include instruction in ways that teachers, principals, pupil services personnel, and school administrators may work more effectively with parents; and

(B) May include activities that:

(i) Involve the forming of partnerships with institutions of higher education to establish school-based teacher training programs that provide prospective teachers and beginning teachers with an opportunity to work under the guidance of experienced teachers and college faculty;

(ii) Create programs to enable paraprofessionals (assisting teachers employed by a local educational agency receiving assistance under part A of title I) to obtain the education necessary for those paraprofessionals to become certified and licensed teachers; and

(iii) Provide follow-up training to teachers who have participated in activities described in subparagraph (A) or another clause of this subparagraph that are designed to ensure that the knowledge and skills learned by the teachers are implemented in the classroom.

"Content knowledge" includes the main ideas, concepts, and syntax of the subject-area domain, the commonly applied algorithms or procedures, and the organizing structures and frameworks that undergird the subject-area domain (Shulman, 1986).

"Pedagogical content knowledge" is an amalgam of knowledge of content and pedagogy that is central to the knowledge needed for teaching. A special kind of professionally useful knowledge of the subject, this knowledge is understanding of "the particular form of content that embodies the aspects of content most germane to its teachability . . [This includes] the most useful forms of representation of those ideas, the most powerful analogies, illustrations, examples, explanations, and demonstrations—in a word, the ways of representing and formulating the subject that make it comprehensible to others. . . . Pedagogical content knowledge also includes an understanding of what makes the learning of specific topics easy or difficult: the conceptions and preconceptions that students of different ages and backgrounds bring with them to the learning of those most frequently taught topics and lessons" (Shulman, 1986, p. 9).

"Curricular knowledge" is an awareness of the full range of programs, texts, and materials designed for the teaching of one's particular topic and grade level as well as a familiarity with the curriculum materials currently used by one's students and their relationships to earlier and later grades' curriculum and with other subjects (Shulman, 1986).

APPENDIX D
LIST OF KEYWORDS USED IN ELECTRONIC SEARCHES

TABLE D1

Professional development keywords used for electronic searches

Keywords	ERIC Thesaurus Term(s)	PsycINFO Thesaurus Term(s)	SocIndex	Professional Development Collection	Dissertation Abstracts
Professional Development	(UT)Professional development; (NT) Faculty development; (R) Staff development; (R) Teacher improvement (R) Inservice teacher education	(UT) Professional development; (R) Inservice teacher education	Use keywords from Keyword Column as needed	Use keywords from Keyword Column as needed	There is an Education, Teacher Training subject category (Descriptor code: 0530) Use keywords from Keywords column as needed
Peer Coaching	(UT) Teacher improvement	Use keywords from Keywords column as needed	Use keywords from Keyword Column as needed	Use keywords from Keywords column as needed	
Teachers' Institutes	(UT) Institutes	Use keywords from Keywords column as needed	(ST) Teachers' institutes	(ST) Teachers' institutes	
Mentoring	(RT) Beginning teacher induction	Use keywords from Keywords column as needed	(ST) Mentoring	(ST) Mentoring	
Teachers' Seminars	(UT) Seminars;	Use keywords from Keywords column as needed	(ST) Seminars; (NT) Workshops	(ST) Seminars	
Teachers' Workshops	(UT) Teacher Workshops	Use keywords from Keywords column as needed	(ST) Teacher workshops	(ST) Teachers' Workshops; (ST) Teacher Centers	

UT: use term

RT: related term

NT: narrower term

BT: broader term

ST: subject term

TABLE D2

Teacher outcomes keywords used for electronic searches

Keywords	ERIC Thesaurus Term(s)	PsycINFO Thesaurus Term(s)	SocIndex	Professional Development Collection	Dissertation Abstracts
Content Knowledge or Curricular Knowledge	Use keywords from Keyword column as needed	Use keywords from Keyword column as needed	Use keywords from Keyword column as needed	Use keywords from Keywords column as needed	There is an Education, Curriculum and Instruction subject category (Use descriptor 0727) Use keywords from Keywords column as needed
Effective Instruction	(UT) Instructional Effectiveness; (R) Program Effectiveness	Use keywords from Keyword column as needed	Use keywords from Keywords column as needed	(ST) Effective Teaching; (RT) Teacher Effectiveness	
Instructional Improvement	(UT) Instructional Improvement; (B) Educational Improvement	Use keywords from Keyword column as needed	Use keywords from Keywords column as needed	(ST) School Improvement Programs; (NT) Curriculum Enrichment	
Instructional Strategies	(UT) Educational Strategies; (R) Teaching Strategies	(UT) Teaching Methods	(ST) Teaching methods	(ST) Instructional Systems; (RT) Teaching	
Pedagogical Content Knowledge	(UT) Pedagogical Content Knowledge; (RT) Knowledge Base for Teaching	(UT) Procedural Knowledge	Use keyword from Keywords column as needed	Use keywords from Keywords column as needed	
Pedagogy	(UT) Instruction; (UT) Teaching Methods;	(UT) Teaching	(ST) Education	(ST) Education; (ST) Logic in Teaching;	
Teacher Attitude	(UT) Teacher Attitudes; (R) Teacher Morale;	(UT) Teacher attitudes; (NT) Teacher expectations;	(ST) Teachers—Attitudes	(ST) Teachers—Attitudes; (NT) Teachers—Attitudes–Evaluation; (NT) Teachers—Attitudes–Research; (ST) Teacher Morale; (R) Teachers—Job Satisfaction	
Teacher Beliefs	Use keyword from Keyword column as needed	(UT) Teacher expectations	Use keywords from Keyword column as needed	(ST) Teachers—Self-Rating of; (ST) Self-efficacy Expectations	
Teacher Change	Use keyword from Keywords column as needed	Use keywords from Keywords column as needed	(ST) Educational Change[1]	Use keywords from Keywords column as needed	
Teacher Self-Efficacy	(UT) Self Efficacy	(UT) Self Efficacy; (R) Academic Self Concept	(ST) Self-Efficacy	(ST) Self-Efficacy	
Teaching Skills	(UT) Teaching Skills; (RT) Teacher Competencies;	Use keywords from Keywords column as needed	(ST) Teaching; (NT) Teaching methods	Use keywords from Keywords column as needed	
Technology Integration	(UT) Technology Integration; (RT) Computer Uses in Education; (RT) Educational Technology	(UT) Instructional Media	(ST) Educational Technology; (NT) Computer -Assisted Instruction; (NT) Computer Managed Instruction	(ST) Educational Technology; (R) Educational Innovations; (R) Teaching – Aids & Devices; (NT) Computer-Assisted instruction	

UT: use term

RT: related term

NT: narrower term

BT: broader term

ST: subject term

TABLE D3

Student achievement keywords used for electronic searches

Keywords	ERIC Thesaurus Term(s)	PsycINFO Thesaurus Term(s)	SocIndex	Professional Development Collection	Dissertation Abstracts
Student Achievement	(UT) Academic Achievement	(UT) Academic Achievement; (NT) Mathematics Achievement; (NT) Science Achievement; (NT) Reading Achievement	(ST) Academic Achievement;	(ST) Academic Achievement;	There is an Education, Tests and Measurements subject category (Use descriptor 0288) Use keywords in the Keywords Column as needed
Student Development	(UT) Student Development; (R) Individual Development	Use keywords in the Keywords Column as needed	Use keywords in the Keywords Column as needed	Use keywords in the Keywords Column as needed	
Learning	Use keywords in the Keywords Column as needed	(UT) Academic Achievement; (B) Learning; (UT) Intellectual Development; (UT) Cognitive Development	(ST) Learning; (NT) Cognitive Learning;	(ST) Cognitive Development; (ST) Learning; (NT) Cognitive Learning	
Student Outcomes	(UT) Outcomes of Education; (RT) Educational Assessment	Educational Measurement	(ST) Educational tests and measurements; (ST) Students--Rating of	(ST) Educational indicators; (RT) Educational accountability	

UT: use term

RT: related term

NT: narrower term

BT: broader term

ST: subject term

TABLE D4
Reading keywords used for electronic searches

Keywords	ERIC Thesaurus Term(s)	PsycINFO Thesaurus Term(s)	SocIndex	Professional Development Collection	Dissertation Abstracts
English	(UT) English, (RT) English curriculum, English instruction	(UT) English, (RT) English as Second Language	Use keywords from Keyword column as needed	(ST) English (RT) English Language—Study and Teaching	There are Language and Literature (Descriptor code: 0279), Reading (Descriptor code: 0535), Education-Bilingual and Multicultural (Descriptor code: 0282) subject categories Use keywords in the Keywords column as needed
Language Arts	(UT) Language Arts, (RT) Language Skills, Literature	(RT) Language Arts Education, Language Development	(ST) Language Arts	(ST) Language Arts	
Literacy	(UT) Literacy, (RT) Literacy Education, Reading Skills, Writing Skills	(UT) Literacy, (RT) Language, Literacy Programs, Reading Development	(ST) Literacy, (RT) Reading, Writing	(ST) Literacy, (RT) Reading, Writing	
Reading	(UT) Reading, (RT) Decoding, Language Processing, Reading Ability, Reading Instruction, Reading Programs, Reading Skills	(UT) Reading, (RT) Reading Education,	(ST) Reading, (RT) Reading—phonetic method	(ST) Reading (RT) Literacy, Reading—phonetic method	
Alphabetics	(UT) Alphabetics	Use keywords from Keyword column as needed.	Use keywords from Keywords column as needed.	Use keywords from Keywords column as needed.	
Composition	(UT) Writing	(UT) Writing	Use keywords from Keywords column as needed.	(ST) Grammar, comparative and general—, composition (language arts)	
Comprehension	(UT) Comprehension, (NT) Listening Comprehension, Reading Comprehension	(UT) Comprehension,	(ST) Comprehension,	(ST) Comprehension, (NT) Learning, Reading Comprehension, Listening	
Fluency	(UT) Reading Fluency, Language Fluency,	(UT) Verbal Fluency, (RT) Language Proficiency, Oral Communication,	Use keywords from Keywords column as needed.	(ST) Fluency (Language Learning)	
Grammar	(UT) Grammar, (RT) Sentence Structure,	(UT) Grammar, (NT) Syntax	(ST) Grammar, comparative & general, Intonation (Phonetics) (NT) Morphology, Phonology, Syntax	(ST) Grammar, Comparative & General, Language & Languages-Grammar	
Letter knowledge	Use keywords from Keyword column as needed.	Use keywords from Keyword column as needed	Use keywords from Keyword column as needed	Use keywords from Keywords column as needed	
Phonemic awareness	(UT) Phonemes, (BT) Phonemics	(UT) Phonological awareness	(ST) Phonemics	(ST) Phonemics	
Phonics	(UT) Phonics, (BT) Phonetics,	(UT) Phonics, (RT) Initial Teaching Alphabet, Reading Education	(ST) Reading—phonetic method (BT) Phonetics	(ST) Reading—phonetic method	

(CONTINUED)

TABLE D4 (CONTINUED)

Reading keywords used for electronic searches

Keywords	ERIC Thesaurus Term(s)	PsycINFO Thesaurus Term(s)	SocIndex	Professional Development Collection	Dissertation Abstracts
Phonological awareness	(UT) Reading Skills	(UT) Phonological awareness, (RT) Phonemes, Phonology, Word Recognition	Use keywords from Keywords column as needed.	(ST) Phonological awareness	There are Language and Literature (Descriptor code: 0279), Reading (Descriptor code: 0535), Education-Bilingual and Multicultural (Descriptor code: 0282) subject categories Use keywords in the Keywords column as needed
Print awareness	Use keywords from Keyword column as needed.	Use keywords from Keywords column as needed.	Use keywords from Keywords column as needed.	(ST) Print awareness	
Vocabulary	(UT) Vocabulary, (NT) Basic Vocabulary, (RT) Vocabulary Development, Vocabulary Skills, Verbal Development	(UT) Vocabulary, (RT) Verbal Communication	(ST) Vocabulary, (RT) Language arts	(ST) Vocabulary, (NT) Word recognition (RT) Vocabulary instruction, Vocabulary in language teaching	
Writing	(UT) Writing, Composition (NT) Paragraph Composition, (RT) Writing Ability, Writing Improvement, Writing Instruction, Writing Processes, Writing Skills,	(UT) Writing Skills, (RT) Literacy, Literacy Programs, Written Communication, Verbal Ability	(ST) Writing, (BT) Communication, (RT) Literacy, Literature, Written Communication	(ST) Writing (RT) Literature, Written Communication, (NT) English Language—Writing, (0T) Composition—Language Arts	

UT: use term

RT: related term

NT: narrower term

BT: broader term

ST: subject term

TABLE D5

Mathematics keywords used for electronic searches

Keywords	ERIC Thesaurus Term(s)	PsycINFO Thesaurus Term(s)	SocIndex	Professional Development Collection	Dissertation Abstracts
Mathematics	(UT) Mathematics, (RT) Mathematical Application, Mathematical Concepts, Mathematics Activities, Mathematics Curriculum, Mathematics Education, Mathematics Instruction, Mathematics Skills	(UT) Mathematics, Mathematics (Concepts),	(UT) Mathematics,	(UT) Mathematics,	There is a Mathematics (Descriptor code:0280) subject category Use keywords from Keyword column as needed.
Algebra	(UT) Algebra, (RT) Prealgebra,	(UT) Algebra, Use term Mathematics to access references from 1973 to June 2003	(UT) Algebra	(UT) Algebra, (RT) Mathematical Analysis	
Arithmetic	(UT) Arithmetic, (RT) Number Concepts, Arithmetic Systems,	(UT) Mathematics	Use keywords from Keyword column as needed.	(UT) Arithmetic, (RT) Mathematical Ability	
Computation	(UT) Computation, Mental Computation	Use keywords from Keywords column as needed.	Use keywords from Keywords column as needed.	(UT) Computational Intelligence	
Data analysis	(UT) Data analysis, (RT), Data processing,	Use keywords from Keywords column as needed.	Use keywords from Keywords column as needed.	(UT) Data Analysis	
Functions	(UT) Mathematics	Use keywords from Keywords column as needed.	Use keywords from Keywords column as needed.	(UT) Functions, (RT) Calculus, Mathematical Models, Algebraic Functions	
Geometry	(UT) Geometry, (RT) Geometric Concepts	(UT) Geometry, Use term Mathematics to access references from 1973 to June 2003	Use keywords from Keywords column as needed.	(UT) Geometry	
Graphing	Use keywords from Keyword column as needed.	(UT) Graphical displays,	(UT) Graphic Methods,	Use keywords from Keywords column as needed.	

UT: use term

RT: related term

NT: narrower term

BT: broader term

ST: subject term

TABLE D6

Science keywords used for electronic searches

Keywords	ERIC Thesaurus Term(s)	PsycINFO Thesaurus Term(s)	SocIndex	Professional Development Collection	Dissertation Abstracts
Science	(UT) Sciences; (R) Science Education; (RT) Science Activities; (RT) Science Curriculum	(UT) Sciences; (UT) Science Education	(ST) Science	(ST) Science; (ST) Science—Study and Teaching	There is a Biological Sciences, General Biology subject category (Descriptor code: 0306) Use keywords from Keywords column as needed
Data Interpretation	(UT) Data Interpretation	Use keywords from Keywords column as needed	Use keywords from Keywords column as needed	Use keywords from Keyword column as needed	
Earth Science	(UT) Earth Science; (RT) Space Sciences	Use keywords from Keywords column as needed	(ST) Earth Sciences	(ST) Earth Sciences	
Experiment	(UT) Science Experiments; (RT) Laboratory Experiments; Laboratory Procedures	Use keywords from Keywords column as needed	(ST) Experimental Design	(ST) Experiments; (RT) Experimental Design	
Exploration	Use keyword from keyword column as needed	Use keywords from Keywords column as needed	Use keywords from Keywords column as needed	Use keywords from Keywords column as needed	
Inquiry	(UT) Inquiry;	(RT) Questioning	(ST) Inquiry (Theory of knowledge)	Use keywords from Keywords column as needed	
Investigation	(UT) Investigations; (R) Evaluation Methods	(UT) Experimental Methods	Use keywords from Keywords column as needed	(ST) Investigations	
Laboratories	(UT) Science Laboratories	(UT) Experimental Laboratories	(ST) Laboratories	(ST) Laboratories	
Life Science	(UT) Biological Sciences	(UT) Biology	(ST) Life Sciences	(ST) Life Sciences	
Observation	(UT) Observation;	(UT) Observation Methods	Use keywords from Keywords column as needed	Use keywords from Keywords column as needed	
Physical Science	(UT) Physical Sciences; (RT) Physics	(UT) Physics; (UT) Chemistry	(ST) Physical Sciences	(ST) Physical Sciences	
Scientific Literacy	(UT) Scientific Literacy	Use keyword from Keywords column as needed	(ST) Scientific Knowledge	Use keyword from Keywords column as needed	
Scientific Procedure	Use keyword from keyword column as needed	(UT) Empirical Methods	(ST) Science—Methodolgy	(ST) Science—Methodology	
Scientific Reasoning	(RT) Science Process Skills	(UT) Reasoning; (R) Hypothesis Testing	(ST) Reasoning	(ST) Reasoning	

UT: use term

RT: related term

NT: narrower term

BT: broader term

ST: subject term

APPENDIX E
RELEVANT STUDIES, LISTED BY CODING RESULTS

Initially relevant studies that did not pass stage 1–full screening (n = 105)

Adenika-Morrow, T. J. (1995). *The TEAM Program: Teaching teachers to utilize an interdisciplinary approach to science for urban students.* Unpublished report. (ERIC Document Reproduction Service No. ED388629)

Adey, P. S. (1995). *The effects of a staff development program: The relationship between the level of use of innovative science curriculum activities and student achievement.* London: King's College London, Centre for Educational Studies.

Adey, P. S. (1997, March). *Factors influencing uptake of a large scale curriculum innovation.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Aloiau, E. K. (2002). Enhancing student motivation in an intensive English language program. *Dissertation Abstracts International, 62*(11), 3671A. (UMI No. 3031494)

Alouf, J. L., & Bentley, M. L. (2003, February). *Assessing the impact of inquiry-based science teaching in professional development activities, PK-12.* Paper presented at the annual meeting of the Association of Teacher Educators, Jacksonville, FL.

Anderson, S. A., Barrett, C., Huston, M., Lay, L., Myr, G., Sexton, D., et al. (1992). *A mastery learning experiment.* Yale, MI: Yale Public Schools.

Appalachian Rural Systemic Initiative. (2000). *Appalachian Rural Systemic Initiative (ARSI): Phase 1. Year 5 annual report.* Lexington, KY: Author.

Appleby, E. (2002). *Pretending to literacy—learning literacy through drama: Evaluation report.* Nathan, Queensland, Australia: Griffith University, Centre for Applied Theatre Research.

Barenholz, H., & Tamir, P. (1997). BIGAL: Biology as a bridge to science in developing communities. *Research in Science & Technological Education, 15*(1), 71–83.

Barfield, S. C., & Rhodes, N. C. (1992). *Review of the sixth year of the partial immersion program at Key Elementary School, 1991–92,* Arlington, VA. Washington, DC: Center for Applied Linguistics.

Bedwell, L. E. (1975, March). *The effects of two differing questioning strategies on the achievement and attitudes of elementary pupils.* Paper presented at the annual meeting of the National Association for Research Science Teaching, Los Angeles.

Beglau, M. M. (2005, July). Can technology narrow the black-white achievement gap? *T.H.E. Journal, 32*(12), 13–17.

Bettencourt, E. M., Gall, M. D., & Hull, R. E. (1980, April). *Effects of training teachers in enthusiasm on student achievement and attitudes.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Blank, R. K., Nunnaley, D., Kaufman, M., Porter, A., Smithson, J., Osthoff, E., et al. (2004). *Data on enacted curriculum study: Summary of findings. Experimental design study of effectiveness of DEC professional development model in urban middle schools.* Washington, DC: Council of Chief State School Officers.

Bos, C. S., Mather, N., Narr, R. F., & Babur, N. (1999). Interactive, collaborative professional development in early literacy instruction: Supporting the balancing act. Learning *Disabilities Research & Practice, 14*(4), 227–238.

Briars, D. J., & Resnick, L. B. (2000). *Standards, assessments—and what else? The essential elements of standards-based school improvement* (CSE Technical Report 528). Los Angeles: University of California, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, and Graduate School of Education and Information Studies.

Brown, M. (2002). Researching primary numeracy. In A. D. Cockburn & E. Nardi (Eds.), *Proceedings of the 26th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 1–015 to 011–030). Norwich, England: University of East Anglia.

Byrkit, D. R. (1968). A comparative study concerning the relative effectiveness of televised and aural materials in the inservice training of junior high school mathematics teachers. *Dissertation Abstracts International, 29*(05). (UMI No. 6816357)

Choike, J. R. (2000). Teaching strategies for 'Algebra for All.' *Mathematics Teacher, 93*(7), 556–560.

Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., et al. (1991). Assessment of a problem-centered second-grade mathematics project. *Journal for Research in Mathematics Education, 22*(1), 3–29.

Cohen, D. K., & Hill, H. C. (1998). *Instructional policy and classroom performance: The mathematics reform in California* (CPRE Research Report Series, RR-39). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Cohen, K. A. (1991). A comparative study of reading instruction management for selected third-grade students in an urban school district. *Dissertation Abstracts International, 52*(08), 2872A. (UMI No. 9201506)

Cotayo, A., Villegas, J. J., Baecher, R. E., & Wilets, I. (1986). *Project MAS 1983-84: O.E.A. evaluation section report.* New York: New York City Public Schools, Office of Educational Assessment.

Couldfield-Sloan, M. B., & Ruzicka, M. F. (2005). The effect of teachers' staff development in the use of higher-order questioning strategies on third grade students' rubric science assessment performance. *Planning and Changing, 36*(3 & 4), 157–175.

Debruhl, D. (1993). The effect of training teachers in peer coaching upon student achievement. *Dissertation Abstracts International, 54*(03), 895A. (UMI No. 9319718)

District of Columbia Public Schools, Division of Quality Assurance. (1986). *Improving basic skills in reading and mathematics: Final evaluation report, E.C.I.A. Chapter 2, 1985–86.* Washington, DC: Author.

Dresner, M. (2002). Teachers in the woods: Monitoring forest biodiversity. *Journal of Environmental Education, 34*(1), 26–31.

Estrada, P. (2005). The courage to grow: A researcher and teacher linking professional development with small-group reading instruction and student achievement. *Research in the Teaching of English, 39*(4), 320–364.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27*(4), 403–434.

Fishman, B. 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*(6), 643–658.

Freeman, D., & Johnson, K. E. (2005). Toward linking teacher knowledge and student learning. In D. J. Tedick (Ed.), *Second language teacher education: International perspectives* (pp. 73–95). Mahwah, NJ: Lawrence Erlbaum Associates.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1982). *Effects of frequent curriculum-based measurement and evaluation on student achievement and knowledge of performance: An experimental study* (Institute for Research on Learning Disabilities Research Rep. No. 96). Minneapolis, MN: University of Minnesota, Institute for Research on Learning Disabilities.

Good, T. L., & Grouws, D. A. (1979). The Missouri mathematics effectiveness project: An experimental study in fourth-grade classrooms. *Journal of Educational Psychology, 71*(3), 355–362.

Good, T. L., Grouws, D. A., & Ebmeier, H. (1983). Elementary school: Experiment II. In T. L. Good,

D. A. Grouws, & H. Ebmeier (Eds.), *Active mathematics teaching* (pp. 93–108). New York: Longman, Inc.

Good, T. L., Grouws, D. A., & Ebmeier, H. (1983). Experimental work in junior high classes. In T. L. Good, D. A. Grouws, & H. Ebmeier (Eds.), *Active mathematics teaching* (pp. 109–141). New York: Longman, Inc.

Harwell, M., D'Amico, L., Stein, M. K., & Gatti, G. (2000, April). *The effects of teachers' professional development on student achievement in community school district #2.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Haughey, M., Snart, F., & da Costa, J. (2001). Literacy achievement in small grade 1 classes in high-poverty environments. *Canadian Journal of Education, 26*(3), 301–320.

Hestenes, D. (2000). *Findings of the modeling workshop project (1994–2000).* Tempe, AZ: University of Arizona.

Hough, D. L. (1994, April). *PATTERNS: A study of the effects of integrated curricula on young adolescent problem-solving ability.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Houston Independent School District, Department of Research and Evaluation. (1996). *Ciencias en Español, 1995-96 (Sciences in Spanish, 1995-96): Research report on educational grants.* Houston, TX: Author.

Hranitz, J. R., & Shanoski, L. A. (1994, October). *Project Success: Challenging children, teachers, and parents to excel.* Paper presented at the Quest for Excellence: Many Paths, Many Voices, One Goal conference, Charleston, SC.

Huffman, D., Thomas, K., & Lawrenz, F. (2003). Relationship between professional development, teachers' instructional practices, and the achievement of students in science and mathematics. *School Science and Mathematics, 103*(8), 378–387.

Huskey, B. (2002). *Academics 2000: Cycle VIII evaluation report, 2001–2002* (AISD Publication No. 01.11). Austin, TX: Austin Independent School District, Office of Program Evaluation.

Irons, J. E., & Carlson, N. L. (2004). Learning styles preparation coupled with teacher assistance: A link to language arts achievement gains for at-risk Hispanic middle school students. In E. M. Guyton & J. R. Dangel (Eds.), *Research linking teacher preparation and student performance: Teacher education yearbook XII* (pp. 233–247). Dubuque, IA: Kendall/Hunt.

Jacob, B. A., & Lefgren, L. (2002). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *The Journal of Human Resources, 39*(1), 50–79.

Jagielski, D. A. (1991). An analysis of student achievement in mathematics as a result of direct and indirect staff development efforts focused on the problem-solving standard of the National Council of Teachers of Mathematics. *Dissertation Abstracts International, 52*(02), 455A. (UMI No. 9119821)

Johanson, G., Martin, R., Gips, C., Beach, B., & Green, S. (1996, April). *The evaluation of the Lead Teacher Project.* Paper presented at the annual meeting of the American Education Research Association, New York.

Johnson, G. L. (2005). The costs and educational effects of the Science Leaders Program: A professional development initiative. *Dissertation Abstracts International, 66*(11), 3986A. (UMI No. 3196209)

Jonesboro School District. (1991). *Incorporating applied learning techniques of basic skills into the secondary vocational education curriculum: Final report.* Jonesboro, AR: Author.

Joyce, B., Calboun, E., Carran, N., & Halliburton, C. (1994, March). *Exploring staff development theories: The Ames study.* Paper presented at the annual conference of the Association for Supervision and Curriculum Development, Chicago.

Kaiser, B. (1996, December). *Teachers, research, and reform: Improving teaching and learning in high school science*

courses. Paper presented at the Global Summit on Science and Science Education, San Francisco.

Kalyani, R., Cohen-Regev, S., & Strobel, S. A. (2001). Student outcomes in a local systemic change project. *School Science and Mathematics, 101*(8), 417–426.

Killion, J. (2003). Use these 6 keys to open doors to literacy: Study of what works by NSOC and NEA distills principles for success. *Journal of Staff Development Council, 24*(2), 10–16.

Kirkwood, M. (2001). The contribution of curriculum development to teachers' professional development: A Scottish case study. *Journal of Curriculum and Supervision, 17*(1), 5–28.

Krol, K., Veenman, S., & Voeten, M. (2002). Toward a more cooperative classroom: Observations of teachers' instructional behavior. *Journal of Classroom Interaction, 37*(2), 37–46.

Lawrenz, F., & McCreath, H. (1988). Integrating quantitative and qualitative evaluation methods to compare two inservice training programs. *Journal of Research in Science Teaching, 25*(5), 397–407.

Levine, D. U., Cooper, E. J., & Hilliard, A., III. (2000). National Urban Alliance Professional Development Model for improving achievement in the context of effective schools research. *The Journal of Negro Education, 69*(4), 305–322.

Lewis, W. H. (1972). The effect of mutual precise goal-setting on teacher- and student-attitudes and on student achievement in elementary science curriculum. *Dissertation Abstracts International, 33*(01). (UMI No. 7220456)

Linek, W. M., Fleener, C., Fazio, M., Raine, I.L., & Klakamp, K. (2003). The impact of shifting from 'how teachers teach' to 'how children learn'. *The Journal of Educational Research, 97*(2), 78–89.

Livesay, M. E., Moore, C. A., Stankay, R. J., Waters, M. J., Waff, D., & Gentile, C. A. (2005). Collaborative learning communities: Building leadership in a high school English department. *English Journal, 95*(2), 16–18.

Luna, E., Gonzalez, S., Robitaille, D., Crespo, S., & Wolfe, R. (1995). Improving the teaching and learning of mathematics in the Dominican Republic. *Journal of Curriculum Studies, 27*(1), 67–79.

Mason, D. A., & Good, T. L. (1993). Effects of two-group and whole-class teaching on regrouped elementary students' mathematics achievement. *American Educational Research Journal, 30*(2), 328–360.

McKenzie, B., & Turbill, J. (1999, November). *Professional development, classroom practice and student outcomes: Exploring the connections in early literacy development*. Paper presented at the joint meeting of the Australian Association for Research in Education and the New Zealand Association for Research in Education, Melbourne, Australia.

Mink, D. V., & Fraser, B. J. (2002, April). *Evaluation of a K-5 mathematics program which integrates children's literature: Classroom environment, achievement and attitudes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Mitchell, B. M. (1994). The ten schools program revisited. *Catalyst for Change, 24*(1), 21–25.

Munoz, M. A. (2002). *Partnership in education: School and community organizations working together to enhance minority students' ability to succeed in high school science and mathematics*. Louisville, KY: Jefferson County Public Schools.

Munro, J. (1999). Learning more about learning improves teacher effectiveness. *School Effectiveness and School Improvement, 10*(2), 151–171.

Nelson, M. A. (1993). The effects of A World in Motion curriculum on selected student outcomes. *Dissertation Abstracts International, 54*(09), 3321A. (UMI No. 9405773)

Nicholson, D. (2006). Putting literature at the heart of the literacy curriculum. *Literacy, 40*(1), 11–21.

Nitsaisook, M., & Anderson, L. W. (1989). An experimental investigation of the effectiveness of inservice teacher

education in Thailand. *Teaching & Teacher Education, 5*(4), 287–302.

O'Connor, R. E. (1999). Teachers learning Ladders to Literacy. *Learning Disabilities Research & Practice, 14*(4), 203–214.

Onchwari, G. (2006). An evaluation of the effectiveness of the national Head Start Bureau early literacy mentor-coach initiative on teacher literacy practices and children's literacy learning outcomes. *Dissertation Abstracts International, 66*(12), 4293A. (UMI No. 3199429)

Osmundson, E., & Herman, J. (2005). *Math and science academy: Year 4 evaluation report* (CSE Tech. Rep. No. 648). Los Angeles: University of California, Center for the Study of Evaluation (CSE)/National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Otto, P. B., & Schuck, R. F. (1983). The effect of a teacher questioning strategy training program on teaching behavior, student achievement, and retention. *Journal of Research in Science Teaching, 20*(6), 521–528.

Pollock, J. (1993). *Final evaluation report, chapter 2: Grade one teacher training program, 1992–93*. Columbus, OH: Columbus Public Schools, Department of Program Evaluation.

Porter, A. C., Blank, R. K., Smithson, J. L., & Osthoff, E. (2005). Place-based randomized field trial to test the effects on instructional practices of a mathematics/science professional development program for teachers. *The Annals of the American Academy of Political and Social Science, 599*(1), 147–175.

Resnick, L. B., Resnick, D. L., & DeStefano, L. (1993). *Cross-scorer and cross-method comparability and distribution of judgments of student math, reading and writing performance: Results from the New Standards Project Big Sky Scoring Conference* (Program two, Project 2.3, CRESST final deliverable). Los Angeles: University of California, Center for Research on Evaluation, Standards and Student Testing.

Reutzel, D. R., Oda, L. K., & Moore, B. H. (1989). Developing print awareness: The effect of three instructional approaches on kindergarteners' print awareness, reading readiness, and word reading. *Journal of Reading Behavior, 21*(3), 197–217.

Rodriguez, A. J., Zozakiewicz, C., & Yerrick, R. (2005). Using prompted praxis to improve teacher professional development in culturally diverse schools. *School Science & Mathematics, 105*(7), 35–362.

Rosebery, A. S., & Warren, B. (2000). *Professional development and children's understanding of force and motion: Assessment results*. Cambridge, MA: TERC, Cheche Konnen Center.

Ross, J. A. (1990). Student achievement effects of the key teacher method of delivering in-service. *Science Teacher Education, 74*(5), 507–516.

Scarborough, J. D. (2004). *Strategic alliance to advance technological education through enhanced mathematics, science, technology, and English education at the secondary level*. Washington, DC: American Association for Higher Education.

Shulman, V., & Armitage, D. (2005). Project Discovery: An urban middle school reform effort. *Education and Urban Society, 37*(4), 371–397.

Shymansky, J. A., Yore, L. D., & Anderson, J. O. (2004). Impact of a school district's science reform effort on the achievement and attitudes of third- and fourth-grade students. *Journal of Research in Science Teaching, 41*(8), 771–790.

Smith, E. L., Blakeslee, T. D., & Anderson, C. W. (1993). Teaching strategies associated with conceptual change learning in science. *Journal of Research in Science Teaching, 30*(2), 111–126.

Snippe, J. (1992, April). *Effects of instructional supervision on pupils' achievement*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Stevahn, L., Johnson, D. W., Johnson, R. T., & Real, D. (1996). The impact of a cooperative or individualistic context on the effectiveness of conflict resolution training. *American Educational Research Journal, 33*(4), 801–823.

Stover, K. W. (2005). Evaluating the effectiveness of teacher training at the Texas Higher Education Collaborative. *Dissertation Abstracts International, 66*(03), 901A. (UMI No. 3170213)

Sunal, D. W. (1991). Rural school science teaching: What affects achievement. *School Science and Mathematics, 91*(5), 202–210.

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2005). The CIERA school change framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly, 40*(1), 40–69.

The Haan Foundation. (2003, March). *Power4Kids, closing the reading gap: Interim report*. Retrieved February 28, 2007, from http://www.haan4kids.org/power4kids/

Tienken, C. H., & Achilles, C. M. (2003). Changing teacher behavior and improving student writing achievement. *Planning and Changing, 34*(3 & 4), 153–168.

Tindal, G., Fuchs, L., Christenson, S., Mirkin, P., & Deno, S. (1981). *The relationship between student achievement and teacher assessment of short- or long-term goals* (Institute for Research on Learning Disabilities Research Rep. No. 61). Minneapolis: Minnesota University, Institute for Research on Learning Disabilities Research.

Togneri, W., & Anderson, S. E. (2003). *Beyond islands of excellence: What districts can do to improve instruction and achievement in all schools--A leadership brief.* Washington, DC: Learning First Alliance.

Van der Sijde, P. C. (1989). The effect of a brief teacher training on student achievement. *Teaching & Teacher Education, 5*(4), 303–314.

Van Haneghan, J. P., Pruet, S. A., & Bamberger, H. J. (2004). Mathematics reform in a minority community: Student outcomes. *Journal of Education for Students Placed at Risk, 9*(2), 189–211.

Van Keer, H., & Verhaeghe, J. P. (2005). Comparing two teacher development programs for innovating reading comprehension instruction with regard to teachers' experiences and student outcomes. *Teaching and Teacher Education, 21*(5), 543–562.

Venville, G., Wallace, J., & Louden, W. (1998). A state-wide change initiative: The Primary Science Teacher-Leader Project. *Research in Science Education, 28*(2), 199–217.

Verville, J. R. (1986). The impact of inservice on secondary teachers receiving content area reading instruction and its effect on student achievement scores: I and II. *Dissertation Abstracts International, 46*(11A). (UMI No. 8522969)

Weinshank, A. B., Polin, R. M., & Wagner, C. C. (1985). *Using student diagnostic information to establish an empirical data base in reading* (Research Series No. 162). East Lansing: Michigan State University, The Institute for Research on Teaching.

Wenglinsky, H. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics*. Princeton, NJ: Educational Testing Service, Policy Information Center, Research Division.

Wenglinsky, H. (2000). *How teaching matters: Bringing the classroom back into discussions of teacher quality*. Princeton, NJ: Educational Testing Service, Policy Information Center.

Westat & Policy Studies Associates. (2001). *The longitudinal evaluation of school change and performance (LESCP) in Title I schools: Final report. Volume 2: Technical report* (U.S. Department of Education, Planning and Evaluation Services Document No. 2001–20). Washington, DC: U.S. Department of Education, Planning and Evaluation Services.

Wilkinson, D., & Luna, N. (1986). *Capital projects, 1985–86: Teach & Reach, Gifted & Talented, BEST*. Austin, TX: Austin Independent School District.

Williams, E. J. (2002). The power of data utilization in bringing about systemic school change. *Mid-Western Educational Researcher, 15*(1), 4–10.

Relevant studies that did not meet stage 2–coding evidence screens (n = 18)

Bahr, C., Kinzer, C. K., & Rieth, H. (1991). An analysis of the effects of teacher training and student grouping on reading comprehension skills among mildly handicapped high school students using computer-assisted instruction. *Journal of Special Education Technology 11*(3), 136–154.

Burkhouse, B., Loftus, M., Sadowski, B., & Buzad, K. (2003). *'Thinking Mathematics' as professional development: Teacher perceptions and student achievement*. Scranton, PA: Author.

Dussault, J. A. (2000). Assessing the effectiveness and implementation of a training module in nonroutine problem solving based on the NCTM's 'Professional Standards' model. *Dissertation Abstracts International, 61*(08), 3060A. (UMI No. 9982796)

Gearhart, M., Saxe, G. B., Seltzer, M., Schlackman, J., Ching, C. C., Nasir, N., et al. (1999). Opportunities to learn fractions in elementary mathematics classrooms. *Journal for Research in Mathematics Education, 30*(3), 286–315.

Howery, B. B. (2001). Teacher technology training: A study of the impact of educational technology on teacher attitude and student achievement. *Dissertation Abstracts International, 62*(03), 861A. (UMI No. 3008753)

Kahle, J. B., Meece, J., & Scantlebury, K. (2000). Urban African-American middle school science students: Does standards-based teaching make a difference? *Journal of Research in Science Teaching, 37*(9), 1019–1041.

Kent, A. M. (2002). An evaluation of the reading comprehension strategies module of the Alabama Reading Initiative with five elementary schools in Southwest Alabama. *Dissertation Abstracts International, 63*(01), 145A. (UMI No. 3040754)

Kowal, P. H. (1989). A study of the effects of selected staff development programs on student learning achievement in mathematics. *Dissertation Abstracts International, 50*(09), 2865A. (UMI No. 9004686)

Lane, M. L. (2003). The effects of staff development on student achievement. *Dissertation Abstracts International, 64*(07), 2451A. (UMI No. 3097871)

MacLean, H. E. (2003). The effects of early intervention on the mathematical achievement of low-performing first-grade students. *Dissertation Abstracts International, 64*(02), 357A. (UMI No. 3081498)

Phillips, G., McNaughton, S., & MacDonald, S. (2004). Managing the mismatch: Enhancing early literacy progress for children with diverse language and cultural identities in mainstream urban schools in New Zealand. *Journal of Educational Psychology, 96*(2), 309–323.

Radford, D. L. (1998). Transferring theory into practice: A model for professional development for science education reform. *Journal of Research in Science Teaching, 35*(1), 73–88.

Scott, L. M. (2005). The effects of science teacher professional development on achievement of third-grade students in an urban school district. *Dissertation Abstracts International, 66*(04), 1268A. (UMI No. 3171980)

Seever, M. (1991). *Summative evaluation of the Comprehensive and Cognitive Development (CCD) program, 1991*. Kansas City, MO: School District of Kansas City, Desegregation Planning Department, Evaluation Office.

Shapley, K. S., Cooter, K. S., & Cooter, R. B. (1999). *Challenges for literacy instruction: The role of teacher capacity building in the Dallas Reading Plan*. Austin, TX: Authors.

Stallings, J., & Krasavage, E. M. (1986). Program implementation and student achievement in a four-year Madeline Hunter follow-through project. *The Elementary School Journal, 87*(2), 117–138.

Villasenor, A., Jr., & Kepner, H. S., Jr. (1993). Arithmetic from a problem-solving perspective: An urban implementation. *Journal for Research in Mathematics Education, 24*(1), 62–69.

Walsh-Cavazos, S. (1994). A study of the effects of a mathematics staff development module on teachers' and

students' achievement. *Dissertation Abstracts International, 56*(01), 165A. (UMI No. 9517241)

Relevant, high quality studies that reached stage 3–detailed coding (n = 9)

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

Cole, D. C. (1992). The effects of a one-year staff development program on the achievement test scores of fourth-grade students. *Dissertation Abstracts International, 53*(06), 1792A. (UMI No. 9232258)

Duffy, G. G., Roehler, L. R., Meloth, M. S., Vavrus, L. G., Book, C., Putnam, J., et al. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects. *Reading Research Quarterly, 21*(3), 237–252.

Marek, E. A., & Methven, S. B. (1991). Effects of the learning cycle upon student and classroom teacher performance. *Journal of Research in Science Teaching, 28*(1), 41–53.

McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, S. N., Cox, S., Potter, N. S., et al. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities, 35*(1), 69–86.

McGill-Franzen, A., Allington, R. L., Yokoi, L., & Brooks, G. (1999). Putting books in the classroom seems necessary but not sufficient. *Journal of Educational Research, 93*(2), 67–74.

Saxe, G. B., Gearhart, M., & Nasir, N. S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. *Journal of Mathematics Teacher Education, 4*(1), 55–79.

Sloan, H. A. (1993). Direct instruction in fourth and fifth grade classrooms. *Dissertation Abstracts International, 54*(08), 2837A. (UMI No. 9334424)

Tienken, C. H. (2003). The effect of staff development in the use of scoring rubrics and reflective questioning strategies on fourth-grade students' narrative writing performance. *Dissertation Abstracts International, 64*(02), 388A. (UMI No. 3081032)

REFERENCES

Ball, D. L., & Cohen, D. K. (1999). Developing practices, developing practitioners: Toward a practice-based theory of professional development. In G. Sykes & L. Darling-Hammonds (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 30–32). San Francisco, CA: Jossey-Bass.

Birman, B., LeFloch, K. C., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., Wayne, A., & Yoon, K. S. (2007). *State and local implementation of the No Child Left Behind Act, volume II—Teacher quality under NCLB: Interim report*. Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

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

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

Clewell, B. C., Campbell, P. B., & Perlman, L. (2004). *Review of evaluation studies of mathematics and science curricula and professional development models.* Submitted to the GE Foundation. Washington, DC: Urban Institute.

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

Cohen, D. K., Raudenbush, S., & Ball, D. L. (2002). Resources, instruction, and research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research*, (pp. 80–119). Washington, DC: Brookings Institution Press.

Cole, D. C. (1992). The effects of a one-year staff development program on the achievement of test scores of fourth-grade students. *Dissertation Abstracts International, 53*(06), 1792A. (UMI No. 9232258)

Corcoran, T. B., Shields, P. M., & Zucker, A. A. (1998). *The SSIs and professional development for teachers*. Menlo Park, CA: SRI International.

Darling-Hammond, L., & McLaughlin, M. W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan, 76*(8), 597–604.

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

Duffy, G. G., Roehler, L. R., Meloth, M. S., Vavrus, L. G., Book, C., Putnam, J., & Wesselman, R. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects. *Reading Research Quarterly, 21*(3), 237–252.

Elmore, R. F. (1997). *Investing in teacher learning: Staff development and instructional improvement in Community School District #2, New York City*. New York, NY: National Commission on Teaching & America's Future.

Fishman, B. 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.

Garet, M., Birman, B. F., Porter, A. C., Desimone, L., Herman, R., & Yoon, K. S. (1999). *Designing effective professional development: Lessons from the Eisenhower program*. Washington, DC: American Institutes for Research.

Garet, M. S., Porter, A. C., Desimone, L. M., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945.

Guskey, T., & Sparks, D. (2004). Linking professional development to improvements in student learning. In E. M. Guyton & J. R. Dangel (Eds.), *Research linking teacher preparation and student performance: Teacher education yearbook XII* (pp. 233–247). Dubuque, IA: Kendall/Hunt.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *The second handbook of research in mathematics education*. Reston, VA: New Age and National Council of Teachers of Mathematics.

Hunter, M. (1984). Knowing, teaching, and supervising. In P.L. Hosford (Ed.), *Using what we know about teaching* (pp. 169–192). Alexandria, VA: Association for Supervision and Curriculum Development.

Joyce, B., & Showers, B. (1995). *Student achievement through staff development: Fundamentals of school renewal* (2nd Ed.). New York: Longman.

Kennedy, M. (1998). *Form and substance of inservice teacher education* (Research Monograph No. 13). Madison, WI: National Institute for Science Education, University of Wisconsin–Madison.

Killion, J. (1999). *What works in the middle: Results-based staff development*. Oxford, OH: National Staff Development Council.

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

Loucks-Horsley, S., Stiles, K., & Hewson, P. (1996). *Principles of effective professional development for mathematics and science education: a synthesis of standards*. Madison, WI: University of Wisconsin at Madison, National Institute for Science Education.

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

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.

Marek, E. A., & Methven, S. B. (1991). Effects of the learning cycle upon student and classroom teacher performance. *Journal of Research on Science in Teaching*, *28*(1), 41–53.

McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, S. N., Cox, S., Potter, N. S., Quiroga, T., & Gray, A. L. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities*, *35*(1), 69–86.

McGill-Franzen, A., Allington, R. L., Yokoi, L., & Brooks, G. (1999). Putting books in the classroom seems necessary but not sufficient. *Journal of Reading Research*, *93*(2), 67–74.

National Commission on Teaching and America's Future (1996). *What matters most: Teaching for America's future*. New York, NY: Author.

National Research Council. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: National Academies Press.

Richardson, V., & Placier, P. (2001). Teacher change. In V. Richardson (Ed.). *Handbook of Research on Teaching* (4th Ed., pp. 905–947) Washington, DC: American Education Research Association.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th Ed.) Thousand Oaks, CA: Sage Publications.

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.

Showers, B., Joyce, B., & Bennett, B. (1987). Synthesis of research on staff development: A framework for future study and a state-of the-art analysis. *Educational Leadership, 45*(3), 77–87.

Sloan, H. A. (1993). Direct instruction in fourth and fifth grade classrooms. *Dissertation Abstracts International*, *54*(08), 2837A. (UMI No. 9334424)

Spillane, J. P. (2000). *District leaders' perceptions of teacher learning* (CPRE Occasional Paper Series OP-05). Philadelphia: Consortium for Policy Research in Education.

Sprinthall, N. A., Reiman, A. J., & Thies-Sprinthall, L. (1996). Teacher professional development. In J. Sikula (Ed.), *Handbook of research on teacher education* (2nd Ed., pp. 666–703). New York, NY: Macmillan.

Supovitz, J. A. (2001). Translating teaching practice into improved student achievement. In S. Fuhrman (Ed.), *National Society for the Study of Education Yearbook*. Chicago, IL: University of Chicago Press.

Tienken, C. H. (2003). The effect of staff development in the use of scoring rubrics and reflective questioning strategies on fourth-grade students' narrative writing performance. *Dissertation Abstracts International*, *64*(02), 388A. (UMI No. 3081032)

U.S. Department of Education (2001). *Teacher preparation and professional development: 2000* (National Center for Education Statistics Report No. 2001-088). Washington, DC: Author.

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.

Yoon, K. S., Garet, M., Birman, B., & Jacobson, R. (2007). *Examining the effects of mathematics and science professional development on teachers' instructional practice: Using professional development activity log.* Washington, DC: Council of Chief State School Officers.