JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 49, NO. 3, PP. 333–362 (2012)

Research Article

Differential Effects of Three Professional Development Models on Teacher Knowledge and Student Achievement in Elementary Science

Joan I. Heller,[1] Kirsten R. Daehler,[2] Nicole Wong,[1]
Mayumi Shinohara,[3] and Luke W. Miratrix[4]

[1]*Heller Research Associates, 230 Grand Avenue, Oakland, California 94610*
[2]*WestEd, 400 Seaport Court, Redwood City, California 94063*
[3]*Teaching and Learning, Vanderbilt University, Nashville, Tennessee 37203*
[4]*Department of Statistics, UC Berkeley, Berkeley, California 94720*

Received 15 June 2011; Accepted 17 December 2011

Abstract: To identify links among professional development, teacher knowledge, practice, and student achievement, researchers have called for study designs that allow causal inferences and that examine relationships among features of interventions and multiple outcomes. In a randomized experiment implemented in six states with over 270 elementary teachers and 7,000 students, this project compared three related but systematically varied teacher interventions—*Teaching Cases*, *Looking at Student Work*, and *Metacognitive Analysis*—along with no-treatment controls. The three courses contained identical science content components, but differed in the ways they incorporated analysis of learner thinking and of teaching, making it possible to measure effects of these features on teacher and student outcomes. Interventions were delivered by staff developers trained to lead the teacher courses in their regions. Each course improved teachers' and students' scores on selected-response science tests well beyond those of controls, and effects were maintained a year later. Student achievement also improved significantly for English language learners in both the study year and follow-up, and treatment effects did not differ based on sex or race/ethnicity. However, only Teaching Cases and Looking at Student Work courses improved the accuracy and completeness of students' written justifications of test answers in the follow-up, and only Teaching Cases had sustained effects on teachers' written justifications. Thus, the content component in common across the three courses had powerful effects on teachers' and students' ability to choose correct test answers, but their ability to explain why answers were correct only improved when the professional development incorporated analysis of student conceptual understandings and implications for instruction; metacognitive analysis of teachers' own learning did not improve student justifications either year. Findings suggest investing in professional development that integrates content learning with analysis of student learning and teaching rather than advanced content or teacher metacognition alone. © 2012 Wiley Periodicals, Inc. J Res Sci Teach 49: 333–362, 2012

Keywords: professional development; science; elementary science; electric circuits; student achievement; teaching cases; looking at student work; metacognition; teacher learning; inservice professional development; content knowledge; English language learners

---

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: National Science Foundation, Teacher Professional Continuum Program; Contract grant number: 0545445.

Correspondence to: Joan I. Heller; E-mail: jheller@edservices.org

DOI 10.1002/tea.21004
Published online 25 January 2012 in Wiley Online Library (wileyonlinelibrary.com).

Conceptual models of effective teacher professional development describe a cascade of influences from features of the professional development to direct impact on teacher knowledge, intermediate impact on classroom instruction, and more distal effects on student achievement (Cohen & Hill, 2000; Desimone, 2009; Heller, Daehler, & Shinohara, 2003; Scher & O'Reilly, 2009; Weiss & Miller, 2006). Although a growing body of literature supports the claim that teacher professional development can improve student achievement (e.g., Blank, de las Alas, & Smith, 2007; Duschl, Schweingruber, & Shouse, 2007; Franke, Carpenter, Levi, & Fennema, 2001; Roth et al., 2011; Saxe, Gearhart, & Nasir, 2001), professional development programs differ widely in the ways they develop teachers expertise and skills (Shulman, 2005; Wilson, Rozelle, & Mikeska, 2010). These variations make it difficult to identify the impact of specific features of professional development interventions on particular aspects of teacher or student outcomes (Fishman, Marx, Best, & Tal, 2003; Scher & O'Reilly, 2009; Wayne, Yoon, Zhu, Cronen, & Garet, 2008).

To build a stronger knowledge base about links among professional development, teacher knowledge, practice, and student achievement, researchers have called for study designs that allow for causal inferences, that isolate treatment effects by systematic comparison of closely related versions of professional development interventions, and that explicitly examine relationships between teacher and student learning (Borko, 2004; Boruch, DeMoya, & Snyder, 2002; Desimone, 2009; Fishman et al., 2003; Jacob, Zhu, & Bloom, 2010; Slavin, 2002; Wayne et al., 2008). Such studies are rare, especially in science. In a review of over 1,300 empirical studies that had the potential to address the link between professional development and teacher learning, only nine met What Works Clearinghouse evidence standards (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). Studies meeting these standards were empirical, employed randomized controlled trials or quasi-experimental designs where groups were matched before the intervention, and included valid measures of student and teacher outcomes. All nine studies focused on elementary school teachers and students. Of those nine, only two focused on science (Marek & Methven, 1991; Sloan, 1993).

Furthermore, the literature to date, including the review by Yoon et al. (2007), largely demonstrates the efficacy of professional development interventions that are delivered by the developers of the professional development courses to relatively small numbers of teachers and schools. A critical step toward scaling up effective practices is to test the delivery of interventions by multiple trainers in a range of typical settings for which the interventions are designed (Borko, 2004; Wayne et al., 2008).

This project was designed to expand the empirical bases for professional development design with a level of rigor that meets the highest evidence standards. Using a randomized experimental design implemented on a large scale in six states, this project compared the differential effects of three related but systematically varied teacher interventions—Teaching Cases, Looking at Student Work, and Metacognitive Analysis—as well as a "business as usual" control condition. The three courses (described in the following section) contained the same subject matter in identical science investigations, but differed in the ways they supported development of teacher pedagogical content knowledge. Interventions were delivered by staff developers trained to lead the inservice courses in their regions, with teacher participants from diverse settings in 39 school districts.

This research used a combination of quantitative and qualitative measures to investigate the impact of each intervention on teacher and student knowledge of the content, on teacher classroom practices, and on teacher pedagogical content knowledge and reasoning about teaching and learning of that content. In addition, the study included systematic collection of observational data capturing participant interactions and reflections both in professional

development sessions and during classroom lessons. In this study, we address the preliminary question of whether the three teacher courses produced teacher and student science learning outcomes that would warrant further analysis of the study's rich set of qualitative data. If so, finer grained analyses of relationships among course features, teacher learning, instructional practices, and student learning will be considered in subsequent papers.

Design of the Professional Development Interventions

Each intervention was based on current beliefs about teacher learning and expertise, and was intended to comprise as strong as possible an exemplar of its kind. All three interventions in this study embodied key features identified in the literature on effective professional development, including: (a) in-depth focus on science content in activities typical of classroom instruction, building on findings that teacher knowledge grows when they encounter subject content through school curricula (Cohen & Hill, 2001; Saxe et al., 2001); (b) opportunities for teachers to engage in active learning; (c) coherence and alignment between the teacher curriculum and standards-based student curricula the teachers were responsible for addressing in their classrooms; (d) substantial duration and length of contact time; and (e) a process of collective participation during which teachers engage in professional discourse and critical reflection (Birman, Desimone, Porter, & Garet, 2000; Desimone, 2009; Yoon et al., 2007).

The three interventions compared were: a *Teaching Cases* course with discussions of prestructured written cases of classroom practice (Barnett-Clarke & Ramirez, 2004; Daehler & Shinohara, 2001); a *Looking at Student Work* course involving analysis of teachers' own student work in conjunction with concurrent teaching (Little, 2004; Little, Gearhart, Curry, & Kafka, 2003); and a *Metacognitive Analysis* course in which teachers engage in metacognitive reflection on their own learning experience (Mundry & Stiles, 2009; White, Frederiksen, & Collins, 2009). Each intervention consisted of 24 hours of contact time, divided into eight 3-hour sessions.

Because the literature contains clear evidence of the critical role that teacher content knowledge plays in raising student achievement (Hill, Rowan, & Ball, 2005; Kanter & Konstantopoulos, 2010), all three interventions included an identical *science content* component that incorporated hands-on science investigations, sense-making discussions, and readings, for half of the course time (Table 1). However, the *pedagogical content knowledge* components were varied to test different approaches to focusing on *learner thinking* and *teaching*.

The science content component was taken from an existing WestEd *Making Sense of SCIENCE* course for elementary teachers on electric circuits (http://www.wested.org/cs/we/view/serv/69). The WestEd course was chosen based on its history of promising effects on elementary student achievement across states, districts of varying sizes, and diverse urban student populations including English language learners (ELL) (Heller et al., 2003; Heller, Daehler, Shinohara, & Kaskowitz, 2004; Heller & Kaskowitz, 2004).

Below, we describe the theoretical underpinnings of the overall professional development approach in all three interventions, for both the science content and pedagogical content knowledge components. We then describe specific features that distinguished the three course configurations and examine the research related to those particular features.

Features of the Science Content Component

Despite the importance of subject matter knowledge, elementary school teachers typically have little training in science and science pedagogy (Fulp, 2002; National Research Council, 2002) and often lack the confidence to teach science (Fulp, 2002; Tosun, 2000). Furthermore,

Table 1
Sources of content and pedagogical content knowledge in three professional development interventions

	Experimental Condition		
Area of Emphasis	Teaching Cases	Looking at Student Work	Metacognitive Analysis
Science content knowledge			
Science investigations	Hands-on, guided investigations to build conceptual understanding		
Discussions	Collaborative sense-making through evidence-based discussion	Same as Teaching Cases	Same as Teaching Cases
Readings	Science content notes and illustrations of classic misconceptions		
Pedagogical content knowledge			
Learner thinking	Analysis of student work and dialog in written cases	Analysis of own students' work from concurrently taught lessons; analysis of assessment tasks	Analysis of teachers' own science learning and thinking
Teaching	Analysis of tradeoffs among instructional options in written cases	Identifying instructional next steps based on evidence of student thinking	Identifying instructional implications of teachers' own learning experience

the content area of electric circuits is known to be particularly challenging for both students (Engelhardt & Beichner, 2004; Shipstone, 1988) and adults (Aschbacher & Alonzo, 2006).

The three intervention courses focused on deepening teacher understanding of core science concepts in national and state standards, leading student curricula, such as Full Option Science for Students (Delta Education, 2010) and Science and Technology for Children (Carolina Curriculum for Science and Math, 2010), Benchmarks for Scientific Literacy (American Association for the Advancement of Science, 1993), and the 2009 National Assessment of Educational Progress (NAEP) Science Framework. Sessions included both grade-level appropriate and more advanced content, such as resistance, to develop teacher knowledge beyond that of their students.

The science content component was designed to immerse teachers in collaborative scientific inquiry to extend their conceptual understanding of key scientific ideas. That is, the purpose of the investigations was to understand phenomena, not to build inquiry skills per se. During *hands-on science investigations*, teachers examined evidence, worked in small groups to make sense of their experiences, and deeply explored their own understandings and misunderstandings. For example, in the first session, groups were provided with a battery, a bulb, and a wire, and challenged to find as many ways as possible to make the bulb light. Then based on this hands-on experience, groups developed their own working definition of a ''complete circuit'' and used their definition to make predictions about other circuits. A facilitated whole-group *sense-making discussion* followed in which teachers shared circuits they built that lit, did not light, and had surprised them. They looked for patterns in their data, and summarized what this helped them understand about circuits. Next, teachers regrouped the data according to complete and incomplete circuits, which predictably led them to discover a

tricky aspect of the science—some complete circuits do not result in a lit bulb (notably, short circuits). To solidify this important understanding, teachers were prompted to describe the relationship between complete, incomplete, lit, and unlit circuits, in a variety of ways and through drawings, writing, and verbal discussion.

During the science content component, course facilitators supported group sense-making by keeping discussions grounded in evidence, prompting teachers to make their thinking visible, and pushing groups beyond surface understandings. Teachers often spent as much time thinking about wrong answers as right answers and gave considerable attention to understanding the thinking behind incorrect mental models and ideas. In addition, the hands-on science investigations were supplemented with substantial *reading materials* that both summarized key science concepts and illustrated common, yet incorrect ways of thinking about the science.

Features of the Pedagogical Content Knowledge Component

While strong subject matter knowledge is essential, it alone is not sufficient for effective teaching. Teachers also need pedagogical content knowledge—subject-specific knowledge about learner thinking and how to teach in a particular discipline (Bransford, Brown, & Cocking, 2000; Shulman, 1986). Teachers with strong content and content-specific pedagogical knowledge have been shown to provide higher quality instruction. More specifically, in the classroom, these teachers are more likely to ask students higher-level questions, use accurate representations and explanations, encourage students to discuss the content and think about applications, and have ideas about and respond to the difficulties students may encounter (Carlsen, 1993; Druva & Anderson, 1983; Hashweh, 1987; Hill & Ball, 2009).

For these reasons, each intervention course also focused on the intersection of knowledge about content and teaching—on developing teacher pedagogical content knowledge. While the approaches in the three courses differed, each was based on the premise that teachers must have opportunities to learn *science content knowledge* in combination with analysis of *learner thinking* about that content and analysis of *teaching* strategies for helping learners understand that content (Shinohara, Daehler, & Heller, 2004; Shymansky & Matthews, 1993; Van Driel, Verloop, & De Vos, 1998).

Honing teachers' abilities to analyze a learner's thinking is key to each of the courses. Such analysis plays an important role in formative assessment, in which teachers' goals are to gather information about student understanding for the purpose of identifying instructional next steps. Prior research has shown that teachers working with colleagues and facilitators in a sustained program to assess student thinking can learn to use evidence from their students' work to revise their teaching strategies, and student performance can improve as a result (Aschbacher & Alonzo, 2006; Gearhart et al., 2006; Gerard, Spitulnik, & Linn, 2010; Kazemi & Franke, 2004; Ruiz-Primo & Furtak, 2007).

Features of the Teaching Cases Course

The Teaching Cases course engaged participants in discussions of narrative cases drawn from actual classroom episodes and written by classroom teachers who work with ethnically, culturally, socioeconomically, and linguistically diverse groups of students. The cases were taken from the same nationally field-tested WestEd *Making Sense of SCIENCE* professional development course that provided the science investigations.

In this course, the cases provided a means of bringing together science, student thinking, and instruction around content-based dilemmas of practice, ones any teacher might face (Barnett-Clarke & Ramirez, 2004; Daehler & Shinohara, 2001; Mundry & Stiles, 2009).

In this way, the course blended an analysis-of-practice approach (Roth et al., 2011) with activities typical of looking-at-student-work professional development experiences (Little, 2004). Analysis of student work embedded in written cases such as those used here may have some benefit over investigations of artifacts from teachers' own classrooms (as in the Looking at Student Work course), in allowing a more sustained focus on the problems of teaching without distraction from teachers' discomfort criticizing each others' practice (Borko, Jacobs, Eiteljorg, & Pittman, 2008; Little & Horn, 2007). Prestructured cases have the potential to support teacher inquiry by making the practice under scrutiny less personal, while also providing an intentional, carefully selected set of student understandings and misconceptions.

The Teaching Cases course was designed to help teachers:

- examine students' science ideas as they pertained to key concepts in electric circuits,
- critically analyze trade-offs among instructional options,
- see content as central and intertwined with pedagogy, and
- focus on the specific content and curricula being taught.

Throughout the course, teachers analyzed cases that contained descriptions of instructional activities, student work samples representing common ways students think about concepts, student-teacher dialogue, and teacher thinking and behaviors. In addition, the hands-on science investigations done by students, as described in the narrative cases, paralleled the science investigations done by teachers in each session. While these cases were not intended as exemplars of best practice, they modeled solid teaching and pedagogical choices known to support student learning.

During each session, teachers worked first in small groups and then as a whole group where they engaged in a subset of the following activities: (a) analyzing the student work presented in a case in terms of correct and incorrect ideas, (b) identifying the logic behind common incorrect science ideas, (c) analyzing the teacher's instructional choices, (d) weighing the tradeoffs of instructional choices in terms of the benefits and limitations of a model, metaphor, definition, or representation used by the teacher in the case, (e) considering the implications for teaching their own students, and (f) reflecting on the process of using cases as a tool for learning.

Features of the Looking at Student Work Course

The Looking at Student Work course engaged teachers in carefully structured, collaborative analysis of their own students' work, which necessarily required that they teach a unit about electric circuits concurrently with participating in the professional development. Compared to the other courses, the Looking at Student Work course directly involved teachers in examining their own students' ideas about electric circuits in the context of their ongoing classroom lessons.

This course utilized artifacts of student work, discussion protocols to keep the attention on evidence of student understanding about circuits, and formative assessments to elicit information about that thinking, components identified as important in the literature (Black, Harrison, Lee, Marshall, & Williams, 2004; Little, 2004; McDonald, 2001). Prior research has shown that teachers can build assessment expertise by working with colleagues and facilitators in a sustained program to analyze student work (Aschbacher & Alonzo, 2006; Gearhart et al., 2006; Gerard et al., 2010), and that student performance improves when teachers use evidence from their students' work to revise their teaching strategies (Gerard

et al., 2010; Ruiz-Primo & Furtak, 2007). Furthermore, in their review of research about teacher science pedagogical content knowledge, Schneider and Plasman (2011) found that teachers who had more experience attending to their students' scientific thinking began to use different assessments in order to gain better information about their teaching.

The Looking at Student Work course was designed to help teachers:

- examine students' science ideas as they pertained to key concepts in electric circuits,
- recognize evidence of incorrect mental models, correct understandings, and proficiency,
- analyze tasks to identify characteristics that support formative assessment, and
- make instructional choices grounded in evidence of student thinking.

In this course, teachers took turns bringing in student work and were given guidance on how to select work samples that best revealed students' science ideas. Each teacher was required to bring in a formative assessment task that they evaluated and refined for use with their own students. Course materials included a task bank of formative assessment items that invited students to write explanations and to draw descriptions of electric circuits phenomena. The richness of the tasks had the potential to provide teachers with data about student understanding that could be analyzed during the course. Teachers could also choose to use the tasks as material resources in their classrooms or as models for the development of similarly informative tasks. This feature of the course design was critical, as Aschbacher and Alonzo (2006) found in their study of elementary school teacher use of science notebooks for formative assessment, teachers need support in developing tasks that are useful for eliciting student understanding.

Throughout the course, the teachers used a written protocol to practice a variety of skills related to analyzing student responses and evaluating the utility of different tasks. During each session, teachers engaged in a subset of the following activities: (a) identified science concepts that were central to a student task, (b) completed the task and analyzed its cognitive demands, (c) identified assessment criteria or constructed an assessment rubric for the task, (d) analyzed the student work in terms of correct and incorrect ideas, as well as common mental models, (e) considered the implications for teaching and learning, (f) described the merits and limitations of the task, and (g) reflected on the process of looking at student work. As with the science investigations, teachers engaged in the analysis of student work via small-group work and whole-group discussion.

Features of the Metacognitive Analysis Course

The third course, Metacognitive Analysis, engaged teachers in reflective discussions about their own learning processes. Rather than analyzing classroom artifacts, such as those incorporated in the other two interventions, the Metacognitive Analysis course utilized teachers' first-hand learning experiences as the objects of analysis of learner thinking. Similarly, teachers' own professional development experiences became the source for reflection on teaching, and eventually the bridge to discussions of implications for their own classrooms and student learning.

Metacognition is an especially powerful tool for adult learning (White et al., 2009) and is linked to interventions that result in greater student achievement in science (Greenleaf et al., 2009).

The Metacognitive Analysis course was designed to help teachers:

- identify concepts that teachers found challenging to learn related to electric circuits,
- examine the logic behind common incorrect ideas pertaining to the topic,

- reflect on their own and others' processes for learning science, and
- analyze the roles of hands-on investigations, discourse, and inquiry in science learning.

In each of the Metacognitive Analysis sessions, teachers engaged in written reflections about their own learning experiences. They were given content-specific prompts that guided their reflections about four areas of inquiry: (a) science ideas they learned during the science investigation, (b) concepts that were particularly tricky or surprising, (c) the logic behind an incorrect science idea that they had, and (d) the implications for what students should learn and how the science content should be taught. The questions for each session corresponded with key concepts presented in that day's lesson. After completing this written reflection, teachers gathered in small groups and then as a whole group to discuss their reflections, revisit areas of conceptual confusion, and briefly discuss implications for their classrooms.

Research Questions

The purpose of this analysis was to investigate whether the professional development interventions improved teacher and student science content knowledge and to compare the differential treatment effects of the three courses. Based on current thinking, improvements in student achievement would not occur without direct professional development effects on teacher knowledge. We therefore included treatment effects on teacher science content knowledge in our analyses.

We assessed science knowledge based on two kinds of evidence: traditional *selected-response tests* of basic factual knowledge, and *written justifications for answers* to selected-response items as measures of richer conceptual understanding. Written justifications of why an answer was selected require the ability to apply science concepts in the service of explaining phenomena, as well as skills in written communication, which are dimensions of content knowledge that are not adequately measured using selected-response items (Quellmalz, Timms, & Buckley, 2005). National Science Education Standards recommend more educational emphasis on assessing rich, well-structured knowledge, as well as scientific understanding and reasoning (National Research Council, 1996, p. 100), and the ability to write detailed explanations is an important part of that ability (Partnership for 21st Century Skills, 2008, p. 4). ''Scientific understanding. . .includes the capacity to reason with knowledge. Discerning what a student knows or how the student reasons is not possible without communication, either verbal or representational'' (National Research Council, 1996, p. 91).

Question 1. What effects do the teacher courses have on teacher science content test scores? If the science investigation component in common among the three courses functioned as intended, during both the study year and follow-up year all three courses would have significant positive effects on teacher science content test scores compared to control teachers, and the course effects would not differ significantly from one another.

Question 2. What effects do the teacher courses have on teacher written justifications? All three courses included both a science investigation component and activities that engaged teachers in writing activities. We expected that during both the study year and follow-up year the three courses would have significant positive effects on teacher written justifications compared to control teachers, and course effects would not differ significantly from one another.

Question 3. What effects do the teacher courses have on student science content test scores? The differences among the three courses should produce different effects on

pedagogical content knowledge and teaching, but in ways that make it difficult to predict student test scores. All three courses were expected to have significant positive effects on student content test scores compared to controls, and no predictions were made as to relative efficacy of the courses.

Question 4. What effects do the teacher courses have on student written justifications? The Looking at Student Work course consisted almost entirely of teachers eliciting and analyzing their own students' written work for evidence of conceptual understanding, as well as analyzing student assessment tasks to identify features that would elicit student thinking. Teachers in this course were also provided a set of written tasks that teachers were encouraged to use with their students and share in the course discussions. This strong emphasis on procedures and materials for looking at student work, combined with students' direct practice of written explanations to provide their teachers with samples of work, were expected to result in significantly stronger written justifications for the Looking at Student Work group than all other groups during the study year. Teaching Cases included analysis of student written work for instructional implications, so would be expected to improve students' written justifications as well compared to the controls in the study year, but would not have as strong effects as the Looking at Student Work condition. During the follow-up year, however, students of teachers who took the Looking at Student Work course in the previous year would not necessarily benefit from direct practice of written explanations unless their teachers again used the task banks as much as they had during the course. We expected that in the follow-up year Looking at Student Work and Teaching Cases would significantly improve students' written justifications compared to controls, but no predictions were made as to relative efficacy of those two courses. Metacognitive Analysis was not expected to improve students' written justifications compared to the controls in either year.

Question 5. What effects do the teacher courses have on English language learner science content test scores? All three courses were expected to have significant positive effects on ELL student content test scores compared to controls. Each course provided first-hand experiences for teachers in ways of learning science that research suggests are particularly well adapted for English learners and other student populations that are severely underserved with respect to science instruction (Lee, 2002). The courses capitalize on the fact that English learners can benefit greatly from inquiry-based science instruction (Hewson, Kahle, Scantlebury, & Davis, 2001); hands-on activities based on natural phenomena depend less on mastery of English than do decontextualized textbooks or direct instruction by teachers (Lee, 2002); and collaborative, small-group work provides opportunities for developing English proficiency in the context of authentic communication about science knowledge (Lee & Fradd, 2001).

Question 6. What effects do the teacher courses have on English language learner written justifications? Following the same reasoning for Question 4, the Looking at Student Work course would have the strongest effects on ELL written justifications, followed in strength by Teaching Cases. However, the challenges of writing science explanations at the fourth grade are even greater for students learning English, and no predictions were made as to treatment effects.

Research Design and Methods

Design Overview

Using a teacher-level randomized trial design, the study compared science content outcomes for three intervention groups and a control group during a study year in 2007–2008,

Table 2
Pretest–posttest follow-up experimental design with four conditions

Group	Random-ization	Fall 2007	Winter 2008	Spring 2008	Summer 2008	Follow-Up Fall 2008[a]
Round 1[b]						
Teachers						
Teaching Cases	R	S–T^T–PD	S–T^T			S–T^T
Looking at Student Work	R	S–T^T–PD	S–T^T			S–T^T
Metacognitive Analysis	R	S–T^T–PD	S–T^T			S–T^T
Control	R	S–T^T	S–T^T		PD	
Students						
Teaching Cases	NR		T^S–Unit–T^S			T^S–Unit–T^S
Looking at Student Work	NR	T^S–Unit–T^S				T^S–Unit–T^S
Metacognitive Analysis	NR		T^S–Unit–T^S			T^S–Unit–T^S
Control	NR		T^S–Unit–T^S			
Round 2[b]						
Teachers						
Teaching Cases	R		S–T^T–PD	S–T^T		S–T^T
Looking at Student Work	R		S–T^T–PD	S–T^T		S–T^T
Metacognitive Analysis	R		S–T^T–PD	S–T^T		S–T^T
Control	R		S–T^T	S–T^T	PD	
Students						
Teaching Cases	NR			T^S–Unit–T^S		T^S–Unit–T^S
Looking at Student Work	NR		T^S–Unit–T^S			T^S–Unit–T^S
Metacognitive Analysis	NR			T^S–Unit–T^S		T^S–Unit–T^S
Control	NR			T^S–Unit–T^S		

R, randomly assigned; NR, not randomly assigned; S, Teacher survey; T^T, teacher content test; T^S, student content test; PD, professional development; Unit, classroom electric circuits unit.

[a]Teachers in the follow-up study were a subset of those in the study year. Students in the follow-up were those in the teachers' classes at that time.

[b]Different cohorts of teachers participated in Rounds 1 and 2.

and delayed effects in a follow-up year. The experimental design in Table 2 shows the data collection events and interventions for teachers, students, and classrooms for the two school years. Intervention teachers were expected to teach all of their classroom electric circuits lessons after they completed the professional development course except for the Looking at Student Work condition that necessitated concurrent classroom teaching. The design included two rounds of interventions and data collection during the study year, and only data collection in the follow-up year. Teachers served as the unit of randomization, and students were nested within teachers. Teachers were randomly assigned to an intervention or control condition and remained in their assigned condition until the conclusion of the study.

The trial was conducted nationally with courses implemented by local facilitators trained at WestEd in Oakland, California. At each of eight research sites during the first year, one or two of the three teacher courses were implemented during each of two rounds, involving different cohorts of teachers. During the follow-up data collection in the next school year, participants included teachers in the three professional development courses and their cohorts of students that year.

Two local facilitators at each research site co-delivered each course, with the exception of the San Francisco Bay Area where WestEd course developers served as solo facilitators for each course. Each facilitator pair taught a different course in Round 1 than they did in Round 2 to avoid confounding facilitator and course effects (Table 3). Having each facilitator pair

Table 3
Counterbalanced research design with three interventions, with teacher samples at random assignment, modified to accommodate site logistics

		Round 1		Round 2				
Site	Facilitator Pair	Summer 2007	Fall 2007	Winter 2007/08	Summer 2008	Intervention Teachers	Control Teachers	Total
1	1	—	CASES	LASW	—	30	10	40
2	2	—	LASW	META	—	27	10	37
3	3	—	META	—	—	15	10	25
4	4	—	CASES	LASW	CASES	45	10	55
5	5	—	LASW	—	CASES	32	11	43
	6	—	META	LASW	—	27	11	38
6	7	—	META	CASES	—	28	10	38
	8	—	LASW	META	—	26	11	37
7	9	CASES	—	META	—	27	10	37
	10	CASES	—	LASW	—	29	10	39
8	11	—	CASES	—	META	17	9	25
	12	—	META	LASW	—	21	10	31
Total						324	122	446

CASES, Teaching Cases; LASW, Looking at Student Work; META, Metacognitive Analysis; —, no course offered.

teach multiple courses introduced the possibility of contamination across conditions from facilitators blurring distinctions among the three experimental interventions. We controlled for such effects with a counterbalanced design such that (a) it included all possible sequences of two courses per facilitator pair over the two rounds; (b) there were overlapping assignments of facilitators so that each course was taught by more than one facilitator pair; and (c) each course variant was taught at least three times in each round. The counterbalanced design allows analysis of both facilitator and treatment effects, without confounding the two. The design controlled for facilitator main effects by having facilitators teach more than one course variant, and if there were order effects from facilitators having previously taught a different version of the course, they would be controlled through systematic variation in course sequences.

While the intended design was perfectly counterbalanced, school and district logistics intervened and the design had to be modified to that shown in Table 3. The modified design maintained the original balance of course offerings, with each course being offered eight times during the study year and 12 courses offered per round of the study. However, three sites needed to teach some courses during the summers before and after the study year instead of running the courses during the school year, and the sequence of courses was reversed at one site.

Research Sites

Regional research sites were identified through a series of discussions with district and county science educators in the United States. The number of fourth grade teachers that were needed for the study restricted the search to large urban districts or to geographic regions consisting of many districts with a smaller number of schools per district. Criteria for research sites included a well-established, stable district or regional science program; strong science leadership (e.g., staff developers, teacher leaders, and district staff) from whom to

draw local course facilitators, and an academically, culturally, and linguistically diverse student population.

The eight research sites that were established through this process included four in the western United States (Arizona, two in California, and Washington), and four in eastern states (Massachusetts, two in North Carolina, and Alabama). Site coordinators were hired to oversee study activities in each region, including recruiting teachers, arranging for meeting and course facilities, running local meetings at which they collected data, pursuing missing data as needed, and supporting local course facilitators and research staff with logistics.

Recruitment of Teachers

Because the topic of electric circuits appears in standards primarily at the fourth grade level, teachers at this grade were invited to participate. Statistical power estimates determined that a sample of 256 teachers (64 per condition) would provide 80% or higher power to detect a minimum effect size (ES) of 0.20 (0.23 for ELLs) at the student level and 0.51 at the teacher level (for $p = .05$). The number of teachers at any particular district in a region depended entirely on teacher interest, as participation was voluntary. Teachers were considered eligible to participate if they had at least 1 year of teaching experience, had not participated in previous *Making Sense of SCIENCE* courses, were teaching fourth grade in the 2007–2008 school year, and expected to do so again in 2008–2009. The teachers received a $650 stipend plus additional stipends if they participated in intensive or follow-up data collection. Students were not randomly assigned, but rather were the students in participating teachers' classes. Active parental consent was solicited through a letter and consent form.

Random Assignment Procedure

Teacher applicants were randomly assigned to one of the experimental conditions using both within-school and between-schools procedures: (a) teachers from schools with two or more participating teachers were randomly assigned within schools, but (b) teachers who were solo participants from their schools were randomly assigned using constructed strata as a blocking factor for teachers. The solo teachers were ranked based on their 2006 school-level state test scores in math, and the ranked list was separated into strata consisting of eight teachers each (or fewer, for odd numbers of teachers). This procedure was followed within each regional site. A total of 446 teachers were randomly assigned to groups (324 to three intervention groups and 122 to control). A randomly selected half of the intervention teachers were then assigned to participate in follow-up data collection in the 2008–2009 school year. Control teachers were not included in the follow-up study because the project had agreed to provide them with professional development courses by the end of summer 2008.

Data Collection

Data were collected before and after two rounds of professional development course implementations from August–December 2007 and January–June 2008, and in the follow-up year. Key teacher and student outcomes reported here include content knowledge in electric circuits, measured by selected-response test items, and quality of written justifications of selected-response answers.

Science Content Assessment. Teacher and student content knowledge about electric circuits was assessed using two tests that were developed and validated in previous evaluations of the WestEd electric circuits course (Heller & Kaskowitz, 2004). The tests were designed to measure a *Making Sense of SCIENCE* content framework that was aligned with National

Science Education Standards Benchmarks, and Full Option Science System (FOSS; Delta Education, 2010) and Science and Technology Concepts (STC; Carolina Curriculum for Science and Math, 2010) curricula. Test questions reflect the format and content of tasks in these curricula and the Trends in International Mathematics and Science Study (TIMSS) and NAEP assessments. Selected-response items always included at least one strong distractor based on a common misconception. The 33-question teacher test and 34-item student test had 15 questions in common, with the other items on the teacher test generally higher level in content and complexity than those on the student test. Pretests and posttests were identical forms of each test.

Program and research staff drafted test specifications and questions, and after internal and external reviews, draft tests were tried out in a series of cognitive interviews. The tests were administered individually to samples of students and teachers drawn from the target populations. The instruments were revised to address identified problems with terminology, representations, and response options, and tested again with additional respondents. The tests were then used in pilot and national field test studies in which they were administered before and after teachers completed WestEd professional development courses from March 2000 through December 2005, and were then modified as needed based on psychometric characteristics.

The teacher test consisted of 20 selected-response items with four or five answer options; 9 yes/no questions, 2 of which also included a constructed-response justification of the answers selected; and 2 additional constructed-response questions involving drawing a circuit or computation. In terms of Webb's depth of knowledge (DOK) levels of cognitive complexity (Webb, 1997; Webb, Alt, Ely, Cormier, & Vesperman, 2006), 24% of the questions were level 1 items involving recall and identification, such as naming the kind of circuit shown in a drawing, or drawing a short circuit; 61% were level 2 items requiring reasoning to predict and describe behavior of circuits, such as determining whether a bulb will light or the relative brightness of bulbs in circuits; and 15% were level 3 items involving application of concepts to justify claims about more complex behavior of circuits, such as explaining why the brightness of a bulb in a parallel circuit is not changed by removing another bulb. The test was designed to be completed in 50–60 minutes. Cronbach's alpha coefficient for the teacher test based on the current study's data was .90.

The student test consisted of 16 selected-response items with four or five answer options; 14 yes/no questions, 3 of which also included a constructed-response justification of the answers selected; and 1 additional constructed-response question involving drawing a circuit. In terms of Webb's DOK levels, 32% of the questions were level 1 items, 56% were level 2, and 12% were level 3. The test was intended to be completed in 30–45 minutes. Cronbach's alpha coefficient for the student test based on the current study's data was .87.

Written Justifications. Teachers and students were asked to write the reasons for their answers to a small number of selected-response items on the content tests to further assess their conceptual understanding and ability to communicate scientific reasoning in writing. For example, student tests included (a) given a drawing of a battery, wire, and bulb showing a short circuit through the jacket of the bulb, ''Will the bulb light?'' and ''Explain why you think the bulb will or will not light'' (student item 12, shown in Figure 1). Both teacher and student tests included (a) a drawing of a battery, wire, and bulb showing a short circuit that did not include the bulb, and asked both ''Is the circuit complete?'' (''Yes'' or ''No'') and ''Explain why you think the circuit is or is not complete'' (teacher item 19 and student item 14); and (b) given a drawing of a parallel circuit with one of its two bulbs missing, ''Will the

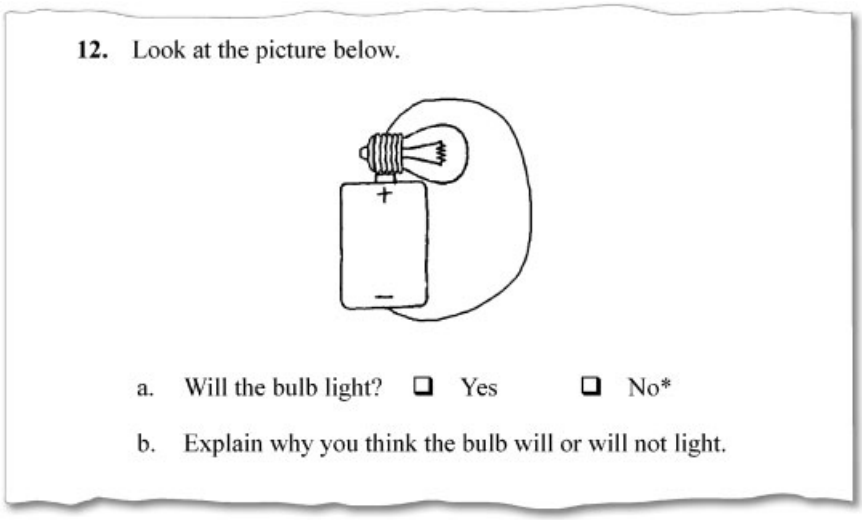

Figure 1. Student written justification test item 12. *Indicates the intended correct response.

bulb light?'' and ''Explain why you think the bulb will or will not light'' (teacher item 20 and student item 15).

Written responses to justification items were scored using item-specific rubrics like the sample rubric provided in the Supporting Information (Scoring Rubric Supplement). The scoring criteria were designed to assess the clarity, precision, accuracy, and completeness of students' written responses. These items and rubrics were reviewed by two science assessment experts who judged that (a) the items and rubrics are typical of standard practice for constructed-response answer justification measures, (b) the rubric scales measure increasingly comprehensive, precise, and explicit understanding of science concepts as well as the ability to write about it in a way that raters can comprehend, and (c) the detailed and item-specific criteria embedded in the scoring rubrics would provide excellent guidelines for scorers and would support reliable scoring compared to the informal coding notes that scorers typically make for themselves during training sessions.

A few examples of student work on item 12 in Figure 1 illustrate the dimensions along which responses varied and were scored. In the drawing, the bulb lies on its side on the positive end of a battery, and a wire goes from the negative end of the battery to the metal jacket of the bulb. Because of the architecture of a typical incandescent bulb, in order for the current to go through the bulb there must be a complete path from one end of the battery to the other end that touches the two contact points of the bulb, the tip and the jacket. The drawing does show a complete, short circuit with current flowing through the metal jacket but not into and out of the bulb. The following student responses would be scored at different levels on the rubric:

> Minimal (0.5 point): ''I think it will not light because it is not put on right.''
> Adequate (1 point): ''The bulb won't light up because the string is pointing to the lumps of the light when it should be pointing at the tip.''
> Thorough (1.5 point): ''I have three reasons for #12. 1st Because the wire is not touching the tip. 2nd Because it is a short circuit. 3rd Because the bulb is not getting energy through the filament.''

Teacher Surveys. An online teaching background survey collected data on all teachers' professional experience and backgrounds in science and teaching as well as school setting and curricula in use. A science teaching survey administered at the beginning and end of the study year and end of the follow-up year elicited a range of self-reported beliefs and classroom practices in relation to science, science teaching, and children's learning. Course evaluation surveys given during the last session of each course measured the degree and quality of the implementation of interventions, and teachers' ratings of the value and impact of the course.

Demographic Information About Students. Teachers provided demographic information about each student taking the test, including sex, race/ethnicity, and English language proficiency category from among (a) Not English proficient—very little or no English; (b) Intermediate English proficient—enough English to participate in classroom interactions; (c) Fully English proficient—home or primary language not English; or (d) Fully English proficient—native English speaker.

Data Collection Procedures

The measures were administered before any teachers in a given round had taken one of the courses, and after the teachers had finished teaching about electric circuits during the school year. Student content tests were administered before and after the electric circuits unit was taught in classrooms of all participating teachers.

Administration of Student Tests of Electric Circuits. A student-testing packet was sent to each participating teacher to provide instructions about how to administer the tests and secure completed tests in sealed envelopes to be returned to the research team. Teachers administered the science tests to their own students, following a detailed testing protocol provided by the research team. Teachers administered pretests during class time within 2 weeks before the electric circuits unit and posttests within 2 weeks after completion of the unit, whenever that occurred during the school year. Students who missed a test because they were absent were given a make-up test as soon as they returned to school. Teachers also completed a classroom and student information survey about student demographics.

Administration of Teacher Tests and Surveys. Intervention group teacher surveys and content tests were administered during the first and last sessions of each teacher course. Control group data were collected by site coordinators in regional project meetings in fall 2007 and winter/spring 2008, after teachers completed electric circuits units in their classes and students had taken their posttests. Site coordinators were provided with detailed test administration instructions to standardize procedures across research sites.

Facilitator Characteristics and Training

Site coordinators and district staff at each site helped identify and solicit the participation of professional development professionals and teacher leaders who might facilitate the courses. The 20 facilitators' classroom experience ranged from 5 to 39 years (mean 19.3 years; median 20 years), and 8 teachers retained teaching responsibilities while serving as facilitators for the study. Facilitators' experience was heavily concentrated in elementary schools, and approximately two-thirds of them had taught at the fourth grade level. Every site had at least one facilitator with fourth grade experience and experience teaching electric circuits. Years of experience as professional developers ranged from <1 to 20 years of experience (mean 6.3 years; median 4 years). As measured by the teacher content test at the end of

the Round 1 training, all facilitators met a reasonable threshold of content knowledge prior to the first round of professional development implementation. Aggregate results showed the mean correct to be 88.3% ranging from 77.8% to 97.8%. There were no significant differences in facilitator content scores by experimental condition, with means of 90.6%, 86.5%, and 87.9% correct for facilitators of Teaching Cases, Looking at Student Work, and Metacognitive Analysis, respectively.

To control threats to implementation fidelity, facilitators initially were trained only on the course they were leading in Round 1. In a 5-day facilitator orientation and training held prior to Round 1 (July 2007), the 10 newly recruited facilitators were told in general terms about the research goals and the three experimental professional development models, and were told which sequence of courses they were assigned to teach. Those who were about to facilitate the Looking at Student Work and Metacognitive Analysis courses learned, for example, that one of the courses included a case discussion component, but in the training they did not read or work with the cases. They were then trained separately in the components of the course to which they had been assigned for the first round.

The approach to facilitator training was analogous to the professional development model in its engagement of facilitators in the (teacher) learners' role. Facilitators experienced the professional development intervention themselves, completing two electric circuits course sessions over 3 days. The majority of the training time was spent deepening facilitators' understanding of electric circuits, grounding them in the common yet incorrect ideas students (and adults) have about the science, and helping participants develop the necessary facilitation skills through observation and practice. Project staff modeled facilitation, engaged the group in analyzing video clips of exemplary facilitation, and provided the trainees with practice in facilitating at least one course session. Facilitators used the course materials throughout the training. After the first round, a second training was held (December 2007) to prepare facilitators for the next courses they would be teaching. Researchers observed and videotaped the trainings to document how the three courses were presented.

Course Implementations

The three professional development courses were delivered eight times each during the first year of the study for a total of 24 course implementations during the study. A total of 283 teachers participated in the first study events held: 201 intervention teachers in the professional development courses, and 82 control teachers in project orientation and data collection meetings.

Each of the three interventions was a 24-hour electric circuits course. Most courses were conducted during the school year with three-hour sessions every other week for 14 weeks. Because of regional logistical issues, however, a small number of courses took place during the summers over one 5-day week.

Two facilitators for each course alternated between serving as lead facilitator and serving as co-facilitator for each session. The materials for each course included a *Teacher Book* that presented all the materials teachers needed to participate in the course, such as the science investigation and group discussion handouts, written cases where relevant, and content guides summarizing the key concepts and outcomes for each session. A *Facilitator Guide* clearly delineated and illustrated the features of each session. It provided detailed yet flexible procedures for leading the course, in-depth background information (e.g., descriptions of the underlying science and common misconceptions), guiding questions and wall charts for each whole-group discussion, and other tips for leading a successful course. In phone debriefs with facilitators before their first session and between subsequent sessions, project staff supported

fidelity of implementation by checking on the content, structure, and process of sessions and reminding facilitators to implement upcoming sessions according to the intended course features.

At the time of the intervention, some teachers were no longer eligible to take the course, either because their school or district did not agree to participate in the study or because they had left teaching or moved to a different grade or school. On average, just over 60% of teachers initially assigned to each intervention group received the intervention, ranging from 40% to 75% at individual sites.

Attendance records were kept for each session of each course implementation. Overall, attendance rates were strong, with almost 95% of the teachers attending all or all but one of the eight course sessions. The frequencies of missing more than two sessions varied, however, from 3% to 4% for Teaching Cases and Looking at Student Work to over 11% for the Metacognitive Analysis course.

It is important to note that the interventions evaluated are not student curricula but rather teacher courses designed to strengthen teaching in a way that is compatible with whatever student curriculum is already used in the classroom. No materials from the Teaching Cases or Metacognitive Analysis courses were provided for use in teachers' classrooms, although some teachers did adapt activities they completed in the course for student use. In the Looking at Student Work intervention, a task bank of formative assessments was available for use with students, but was optional as teachers were free to use assessments from their own student curriculum when collecting samples of student work.

Analytic Samples

Just over a third of the 446 randomly assigned teachers (156) dropped out of the study before attending any project events or providing any data, generally because of scheduling conflicts or time constraints (see Table S1 in the Supporting Information). The control group had the fewest teachers leaving the study at this stage, not surprisingly since they had much lower likelihood of a time conflict with their brief project meetings than intervention teachers had with their 24-hour courses. Of the 290 teachers remaining, only 19 additional teachers (6.6%) dropped out after attending one or more meetings or course sessions. The Metacognitive Analysis teachers had the highest dropout rate, with 12.5% leaving during the study.

The analytic teacher sample was defined as all teachers with complete data, including pretest, posttest, and demographic/educational background covariates that were included in the hierarchical linear model (HLM) analyses. A sample of 271 fourth grade teachers was retained through the end of the study and provided teacher data sets; 253 of these teachers provided test data from their students. The analytic student sample was defined as all of their students with pretest, posttest, and parent consent. Of the teachers retained in the study as of the end of study, 71 also provided teacher and/or student data in the follow-up year. This follow-up sample corresponds to approximately 25% of the 283 teachers who attended the first event in the study year.

Baseline Equivalence of Samples

The internal validity of the study depends upon baseline equivalence among intervention group and control group teachers and students. As would be expected with teacher-level random assignment, the four groups were comparable with respect to teachers' self-reported educational backgrounds, and teaching and professional development experience. Between 93% and 99% of the teacher sample had bachelor's degrees, and about half also had a

master's degree. The median teaching experience was 7–9 years, with median electric circuits teaching experience of 3 years. Professional development experience in the past 3 years was also comparable, with medians of 20–24 hours in science and 6–8 hours in electric circuits. It is notable that the sample includes teachers with a wide range of teaching and professional development experience, spanning from novice to veteran teachers in all groups.

In every group, the teacher sample ranged from 80%–90% female, 60–70% White, 9–14% Black, 1–10% Hispanic or Latino, and 3–4% Asian. In all ethnic categories there was variation among the four experimental conditions, but no consistent pattern indicated bias. Details on teachers' education, experience, and demographics are provided in the Supporting Information (Teacher Background Supplement).

Teachers' and students' pretest scores on the tests of electric circuits content knowledge were also examined for equivalence. There were no differences among the teacher means, all of which were between 56% and 61%, nor among the student means, which were between 48% and 49%.

Data Analysis Procedures

To address questions about the relative impacts of the interventions, test and written justification data for teachers and students were first scored and then HLM analyses were used to estimate the treatment effects of the interventions on each outcome.

Scoring of Content Knowledge Tests

Percent correct scores were computed for selected-response items on teacher and student content knowledge tests, with each item worth one point. Scores on written justifications were not included in content test scores.

Scoring of Written Justifications

Four raters scored written responses to justification items after training and calibration sessions on the use of rubrics specific to each question. A sample rubric is provided in the Supporting Information (Scoring Rubric Supplement). All responses were double-scored and discrepancies between scores on an item were addressed in one of two ways. First, if the discrepancy was one point or more, it was considered an indication that something was amiss in one or both scorers' interpretations of the rubric or of the written response, and the scorers discussed their scoring rationales in discrepancy resolution sessions with a third rater present. In almost all cases, the discrepancy was resolved to consensus. For the $<1\%$ of discrepancies on which the scorers could not reach consensus, the mean of the two scores was included in the analytic data set. If the discrepancy was a half-point, the mean of the two scores (0.25 above the lower score and below the higher score) was included in the analytic data set. Final scores included in the analyses consisted of the sum of all points on written justification responses (maximum possible score of 5 points for students and 3.5 points for teachers).

Impact Analyses

HLM models were fitted to gain scores from pretest to posttest and, for teachers, from pretest to follow-up test. Separate models analyzed teacher outcomes, student outcomes, and students classified as ELLs. Analyses were conducted separately for teacher gain scores on the selected-response questions on the test and for their written justification responses, and data from the first year of the study were analyzed separately from the follow-up data from the next year. Control teachers did not provide data in the follow-up year but control data from the study year were included in analyses to roughly approximate group differences.

The teacher model was a two-level HLM with teachers nested in professional development course groups, that is, the groups of teachers who took each course together. These course groups were expected to vary beyond differences related to the type of intervention, both as a result of the characteristics of the facilitators and the interactions among the teachers within the class. The student model was a three-level HLM. Students were nested in classes taught by specific teachers nested in the professional development course groups. Details of the analytic models are provided in the Supporting Information (Data Analysis Supplement).

Results: Impacts of Interventions

This study provided strong evidence of efficacy for the three professional development courses tested in that all produced significant increases in teacher and student outcomes. As summarized in Table 4, the interventions all brought about highly significant gains in teachers' and students' scores on selected-response tests of science content knowledge, well beyond those of comparable control groups. The score increases for students of intervention teachers also occurred for ELLs. There were no significant differences among student gains based either on sex or race/ethnicity. Furthermore, the powerful treatment effects were maintained in the school year following the study year, when both intervention teachers and their next cohort of students showed gain scores significantly greater than those of controls (based on the study year control scores).

With respect to the second measure of teacher and student content knowledge, the quality of written justifications, a different pattern was observed. For teacher written justifications, in the study year all three interventions brought about highly significant gains in teachers' scores as compared with controls. However, in the follow-up year, only the treatment effects of the Teaching Cases course again were significantly greater than controls. For students, in the

Table 4
Effect sizes and significance levels for three interventions compared to controls

Measure	Teaching Cases	Looking at Student Work	Metacognitive Analysis
Teacher			
Content knowledge study year	1.84***	1.81***	1.93***
Content knowledge follow-up year	1.04**	1.45***	1.21***
Written justifications study year	0.68**	0.64**	0.58*
Written justifications follow-up year	0.70*	0.34	−0.05
Student			
Content knowledge study year	0.37***	0.57***	0.60***
Content knowledge follow-up year	0.48**	0.50***	0.75***
Written justifications study year	0.01	0.31**	0.07
Written justifications follow-up year	0.39*	0.42*	0.21
English language learner			
Content knowledge study year	0.72***	0.76***	0.73**
Content knowledge follow-up year	1.01**	0.84*	1.33***
Written justifications study year	0.17	0.35	−0.12
Written justifications follow-up year	−0.03	0.51†	0.09

*$p < 0.05$, two-tailed.
**$p < 0.01$, two-tailed.
***$p < 0.001$, two-tailed.
†$p < 0.10$, two-tailed.

study year only the Looking at Student Work course significantly raised written justification scores. In the follow-up year, the delayed effects of the Teaching Cases course matched the effects of the Looking at Student Work course, and both significantly raised answer justification scores. Note that other than teachers' study year results, the Metacognitive Analysis course did not raise teachers' or students' scores on written justifications.

Detailed results of the HLM analyses are provided in Tables S19, S20, and S21 of the Supporting Information, for teachers, the full sample of students, and ELLs, respectively. Unadjusted mean scores for these and all other analyses are provided in the Supporting Information (Unadjusted Results Supplement).

Teacher Results

Question 1. Effects on Teacher Science Content Test Scores. All three interventions caused sizable content test score gains for teachers as shown in Table S19 in the Supporting Information. The three experimental conditions raised teachers' scores on the test of electric circuits knowledge by approximately 22 percentage points from study pretest to posttest. Estimated gains for all three intervention groups were significantly greater than control group teachers' estimated gains of 2.4 percentage points ($ES = 1.8$–1.9). There were no significant differences among the impact estimates of the three courses. Unadjusted mean scores for these and all other analyses are provided in the Supporting Information (Unadjusted Results Supplement).

Furthermore, the intervention teachers' significantly higher content test gains were maintained a year after the professional development, with intervention teachers' estimated gains of 14–18 percentage points higher than the control teachers' gains in the previous year.

Question 2. Effects on Teacher Written Justifications. Teachers' responses to answer justification items followed the same pattern as for the selected-response items, with teachers in all three experimental conditions demonstrating estimated gain scores significantly greater than control group teachers' impact estimates, with impact estimates for the three interventions of approximately 0.6 points from study pretest to posttest versus control group estimated gains of 0.1 points ($ES = 0.6$–0.7; see Table S19 in the Supporting Information). There were no significant differences among the impact estimates of the three courses. With respect to the follow-up, only the Teaching Cases course led to significant effects, with an adjusted gain of 0.75 points ($ES = 0.70$), still far greater than the control group gain in the study year.

Student Results

Question 3. Effects on Student Science Content Test Scores. Students of teachers in all three experimental conditions demonstrated significantly higher estimated gains than control group students, with average gains of 19–22 percentage points compared to 13 percentage points for control students (see Table S20 in the Supporting Information). There were no significant differences among the three interventions in their effects on student content test scores.

Teachers' cohorts of students in the follow-up school year also demonstrated clear benefits from the professional development courses taken the previous year. The interventions improved follow-up students' content test scores from 19 to 23 percentage points, remaining significantly greater than the control group gains of under 13 percentage points in the previous study year. While no differences were found among the three courses in their impact on student content test scores in the year after the interventions, all of them had powerful and sustained impacts compared to the control condition.

Question 4. Effects on Student Written Justifications. Students' written justifications did not follow the same pattern as for the selected-response items. Only the Looking at Student Work course significantly improved scores compared to controls from study pretest to posttest (see Table S20 in the Supporting Information). In the follow-up year, however, students of both Teaching Cases and Looking at Student Work teachers demonstrated gain scores that were significantly higher than the control students' study year gains. Metacognitive Analysis did not improve students' written justifications compared to the controls in either year.

English Language Learner Results

Question 5. Effects on ELL Science Content Test Scores. The findings presented thus far have been for the sample of all students in participating teachers' classrooms. Analysis of ELLs' content test data revealed the same pattern as for the full sample (see Table S21 in the Supporting Information). The three interventions raised ELL students' scores by approximately 18 percentage points, all three significantly greater than control group students' estimated gains of 7.1 percentage points ($ES = 0.72$–0.76). There were no significant differences among student scores based either on sex or race/ethnicity. There were no significant differences among the impact estimates of the three courses.

In the follow-up year the interventions again raised ELL students' test scores by an estimated 19–27 percentage points, significantly more than the control ELL gains in the previous year.

Question 6. Effects on ELL Written Justifications. ELL student written justifications did not show significant treatment effects during the study year. During the follow-up year, the Looking at Student Work course improved ELL students' written justifications by an estimated 0.7 points, marginally more than control group ELL students' written justification gains of approximately 0.3 points ($p < 0.10$, $ES = 0.51$).

Relationships Among Teacher and Student Outcomes

The findings indicate that all three professional development courses increased both student and teacher science content knowledge. To examine the relationships among gains in teacher content knowledge and student achievement, we determined whether teacher posttest content knowledge predicted gains in student content test scores. We estimated the impacts on selected-response item scores with an HLM model for teacher content knowledge without including experimental condition in the model. Results indicated that teacher content knowledge was a significant predictor of student test scores ($p < 0.001$).

To determine whether teacher content knowledge accounts for most of the student outcomes, we compared the results for the HLM model that had only teacher content knowledge to a model that had both teacher content knowledge and the experimental condition dummy variable. A likelihood ratio test indicated that the models were different ($p < 0.01$), and all three treatment effects were significantly positive ($p < 0.05$, $p < 0.005$, $p < 0.005$ for Teaching Cases, Looking at Student Work, and Metacognitive Analysis, respectively). We concluded that all three teacher interventions improved student test scores only in part through their effects on teachers' content knowledge.

Discussion

This study was part of a larger project designed to identify links between features of professional development and outcomes for teachers and students. The project not only incorporated experimental methods that permit causal inferences about the effects of three

systematically varied interventions, but also rich, qualitative measures to illuminate the quantitative results. Data were collected to document the courses as actually enacted, as well as the impact of each intervention on teacher content and pedagogical content knowledge, classroom practices, and student achievement. This study of multiple interventions, each delivered by multiple staff developers in diverse contexts like those for which the professional development was intended, is a rare example of the kind of ''Phase 3 research'' called for by Borko (2004) ''to progress toward the goal of providing high-quality professional development for all teachers'' (p. 4). This paper focused only on the preliminary question of whether the courses improved teacher and student science content knowledge enough to warrant further analysis of the qualitative data collected. Indeed, the three teacher courses all had powerful effects on science learning for both teachers and students, including ELLs, as well as differential effects on teacher and student outcomes.

Impact on Science Content Test Outcomes

Teachers. The three interventions had identical collaborative science activities that engaged teachers in investigating and making sense of elementary grade electric circuits content for half of each course session. All three courses produced large gains in teacher content knowledge in the study year, with effect sizes close to 2.0 and over 1.0 a year later, and there were no significant differences among the course effects on teacher test scores. The science content component, which was drawn from a WestEd *Making Sense of SCIENCE* professional development course, provided a powerful learning experience for teachers. Science content in the teacher courses was presented in the context of the student curricula teachers were using in their classrooms, and previous research has shown that professional development is most successful when there is this kind of alignment between the teacher curriculum and standards-based student curricula (Cohen & Hill, 2001; Desimone, 2009).

Students. All three courses produced significant increases in content test scores compared to controls, both in the study year and for students in treatment teachers' classes a year later. Effect sizes ranged from 0.4 to 0.6 for students in the study year, and were even stronger (0.5–0.8) the following year. To our knowledge, treatment effects of this magnitude and duration have not been reported in previous research. Since the courses greatly improved teacher content knowledge, we tested whether those increases alone produced the large increases in student test scores. Indeed, teacher gains in content knowledge significantly predicted student gains, but the considerable impact of the courses on teacher content knowledge only partially accounted for student outcomes. Teachers get something else out of the courses—perhaps additional pedagogical knowledge that influences their teaching practices. Analyses of data collected in this study on teachers' pedagogical content knowledge, teaching practices, and reasoning about teaching should reveal some of these other treatment effects.

English Language Learners. All three courses had even stronger effects on ELL content test scores than were found for the full sample of students, with effect sizes of 0.7–0.8 in the study year, and 0.8–1.3 for students in treatment teachers' classes a year later. Again, to our knowledge this is the first study to document such strong effects on English learner science achievement. Research has shown that English learners can benefit greatly from inquiry-based science instruction (Hewson et al., 2001); hands-on activities based on natural phenomena depend less on mastery of English than do decontextualized textbooks or direct instruction by teachers (Lee, 2002). In addition, collaborative, small-group work provides opportunities for developing English proficiency in the context of authentic communication about science

knowledge (Lee & Fradd, 2001). Teachers experienced these features in the three courses, which may have influenced their classroom instruction, benefiting English learners as well as the rest of their students.

Impact on Written Justifications

Teachers. Because the three courses all included both a science investigation component and other activities that engaged teachers in writing about their science ideas, we expected the three courses to have significant positive effects on written justifications for selected-response answer choices compared to control teachers, during both the study year and follow-up year. All three courses did significantly raise teacher written justification scores in the study year, but only Teaching Cases had significant effects in the follow-up year. This is an interesting result in that Teaching Cases was the only course that engaged teachers in critical analysis of tradeoffs among instructional options, with detailed consideration of science content embedded in decisions about classroom practice. In this course, teachers also examined a purposeful selection of student work in written cases, thus exposing them to the most common misconceptions learners tend to display. This exposure may have deepened teacher conceptual understanding of the science. In contrast, in the Looking at Student Work course, the set of misconceptions encountered in work from teachers' own students would likely have been less comprehensive.

Students. Students' written justifications did not follow the same pattern as for the selected-response items—the three courses differed with respect to sources of pedagogical knowledge, and their efficacy for improving ability to justify answer choices varied accordingly. Only the Looking at Student Work course significantly improved student written justification scores compared to controls in the study year, when Looking at Student Work teachers taught the unit on electric circuits concurrently with taking the professional development course. This meant that students were completing writing tasks that were assigned as part of their teachers' Looking at Student Work course shortly before the students took the content post-test, giving them a considerable advantage over students in the other two intervention groups.

Furthermore, Looking at Student Work was the only course that had teachers rehearse a classroom practice skill, that of administering assessment tasks to elicit student thinking. It is essential that professional development supports teachers' skill in the instructional routines at the heart of classroom practice (Ball & Cohen, 1999; Dewey, 1965; Grossman, 2005; Grossman & McDonald, 2008; Lampert, 2010); teachers must "learn how to do instruction, not just hear and talk about it" (Ball, Sleep, Boerst, & Bass, 2009, p. 459), and the formative assessment experience may have supported more use of classroom written science explanation tasks.

Interestingly, student gains in written justification scores were not seen for the Teaching Cases course until the follow-up year. In contrast, other than teachers' study year results, the Metacognitive Analysis course did not raise teachers' or students' scores on written justifications. This pattern likely reflects the fact that the Teaching Cases and Looking at Student Work courses both included analysis of student work and attention to the importance of using classroom tasks that elicit useful information about students' conceptual understandings, whereas the Metacognitive Analysis course included similar activities but focused only on the teachers' own learning experience and understandings. Although further research is needed to understand this result, one possibility is that teacher metacognition about their own learning does not necessarily lead teachers to insights about student difficulties or changes in teaching practice.

English Language Learners. ELL student written justifications did not show significant treatment effects during the study year. During the follow-up year, the Looking at Student Work course did improve ELL students' written justifications, but at a marginally significant level ($p < 0.10$). Writing accurate and complete science explanations is very complex and difficult for fourth graders and would be even more difficult for English learners. The fact that there were indications of improvement in the follow-up year likely reflects the strong emphasis on eliciting student thinking in the Looking at Student Work course. Furthermore, the significant score increases for ELL students of intervention teachers, and lack of differences among student gains based on sex or race/ethnicity suggest that all three of the courses have design features that are effective at preparing teachers to support all students' science learning.

Implications of the Findings

This study demonstrated that professional development of moderate duration—in this case, one 24-hour course—can have considerable and lasting impact on teaching and learning in elementary science. With high-quality professional development, it was possible to deliver courses in multiple settings by non-developer facilitators and achieve effect sizes of 2.0 for teachers and 0.7–1.0 for students, with effect sizes highest for English learners. Effects were stronger for intervention teachers' students in the follow-up year, suggesting that course impacts were not fully realized until teachers had time to process and implement what they learned.

Design of Professional Development. While this experiment compared models corresponding to general types of professional development (e.g., case discussions, looking at student work), the results should not be generalized to those broad categories. Interpretation of the findings requires a closer analysis of the relationship between specific features of the courses and teacher and student outcomes. That is, each course had features unique to the particular versions implemented here and it is more important, for example, to recognize the different ways in which all three courses included components related to analyzing student work than to think of the Looking at Student Work course as representing that genre of professional development.

It is notable that all three courses raised teacher and student test scores well beyond those of controls, and the effects were even stronger a year later. It is extremely rare for research to show such a powerful and sustained link between professional development and student achievement, especially in science, and this may indicate that the science investigations in common across the courses should be candidates for widespread implementation and further study. However, the courses with the strongest impacts on written justifications of answers emphasized science content situated in activities and scenarios involving student curricula and instruction, in combination with analysis of student work and classroom pedagogical practice. Based on these findings, policy makers should invest in professional development that emphasizes analysis of student learning, pedagogy, and content, rather than focusing on general pedagogy or purely on content. Furthermore, the beneficial effects of both the Teaching Cases and Looking at Student Work courses may point to the potential of a program that combines both of these approaches.

The results for English learners are also noteworthy. The interventions targeted regular science teachers and classes, not pull-out classes taught by specialists in teaching ELLs. This suggests that the approaches in these courses have the potential to benefit the large majority of ELLs in a way that enhances all students' opportunities to learn.

Dissemination of Professional Development. Unlike previous studies in which developers of professional development delivered courses directly to small groups of teachers under ideal conditions, in this and previous studies of *Making Sense of SCIENCE* courses, the professional development was delivered through cadres of staff developers who were trained to lead teacher courses in their regions. This approach included a combination of leadership academies, written facilitator guides, and opportunities to debrief by phone with project staff as the key mechanisms of support. The positive outcomes indicate that the train-the-trainers model has the potential for broad dissemination and impact at a relatively low cost. While there is a considerable body of research on professional development for teachers, there is almost no research on preparation of facilitators of professional development. The approach to facilitator training and support in this project could provide an opportunity for structured studies of effective facilitator preparation.

Research Directions. This study provided a clear answer to the preliminary question about teacher and student learning, demonstrating that the courses do lead to strong and positive outcomes. From this broad causal connection alone, however, it is not possible to trace the cascade of influences by which the courses achieved such strong outcomes. That is, we know the effects of each course but there are multiple variables at work in each course design. We need finer-grained analyses of the qualitative and observational data to illuminate processes and relationships underlying the quantitative patterns. For example, we speculated that in the Teaching Cases intervention, the intentional selection of misconceptions evidenced in student work in the written cases may have deepened and extended teachers' conceptual understanding of the science, whereas in Looking at Student Work there would be more limited exposure to common but incorrect ways of thinking about the science. Video of professional development sessions can be analyzed to determine how discussions of student work in these two courses compare substantively in scope or depth and to identify conceptual and pedagogical affordances of the differences between the courses. Linkages can then be explored to classroom teaching, teacher pedagogical reasoning interviews, and written responses to student work, to trace connections from professional development through teacher knowledge about student work to student opportunities to communicate their science ideas in those teachers' classrooms. In this way it may be possible to tease out hypotheses as to the influence of different features of the professional development courses on classroom practice, and in turn, student learning.

More generally, although it is often not feasible to do large-scale randomized experiments, research is needed that takes on the challenging task of making connections among the features and processes of professional development, impacts on teacher knowledge, intermediate impact on classroom instruction, and indirect effects on student achievement.

Measures are Crucial. Measures must be used that are sensitive to differences among interventions. In this study, all courses raised selected-response test scores, but written justifications of selected-response answers revealed conceptual understandings and ability to communicate about science that did differentiate among the effects of the courses. For such young students, other measures that depend more on drawings or verbal interactions with students might be used to gain additional information about conceptual understanding and scientific communication at this grade level.

Strengths and Limitations of the Analyses

The power in the design of this study lies in the combination of several elements. The study compared carefully configured professional development designs, with both shared and

differing components. It utilized a set of measures driven by a conceptual logic model of the professional development's target outcomes, and implemented a rigorous randomized experimental design that permitted inferences about causal relationships.

Since this study recruited a volunteer sample, these findings should only be generalized to teachers for whom the tested professional development is a priority. This holds for the original 446 teachers who were recruited and randomly assigned to intervention and control groups, and for the 271 teachers remaining after attrition who provided teacher and/or student data.

Finally, the meaning of this study depends upon the validity of the measures used, and the measure of written justification was extremely difficult for the fourth grade sample. This measure was sensitive enough to detect differences among the professional development interventions, so is promising, but for this age group the cognitive load of writing coherent statements limited the measure's utility for assessing conceptual understanding. As a result, a large proportion of the responses were missing or irrelevant, and the distribution of responses violated some assumptions of a hierarchical linear analysis.

The authors would like to acknowledge colleagues who made the study possible from the early design phases to the final analyses. Judith Warren Little was a co-Principal Investigator in this collaborative project between Heller Research Associates, WestEd, and University of California, Berkeley, and made invaluable contributions at all stages of the study.

We thank the site coordinators who made the program implementation and teacher data collection possible: Kim Bess, Marlene Kotelman, Cindy Moss, Carol Mueller, Darlene Ryan, Andrew Schwebke, and Meg Watson. We are also grateful to the numerous course facilitators who so skillfully delivered the professional development courses, working diligently to stay true to the specifications of each course in the sequence they were assigned. The participating teachers and their students around the country contributed the data on which this entire study depended. We recognize the burden associated with taking part in a research study of this magnitude and thank them for their time, commitment, diligence, and interest over the past several years.

We are grateful to Paul LeMahieu, Sophia Rabe-Hesketh, Ed Haertel, and Karen Sheingold for crucial advice at key points in the research design and data analyses. Karen also contributed in major ways to our pedagogical content knowledge framework and by providing input on drafts of this manuscript. This study could not have been done without the unwavering commitment of the implementation team and the entire staff that supported the project: Michelle Simone, Jennifer Mendenhall, and Mikiya Matsuda. We extend huge thanks to Cara Peterman Price at Heller Research Associates for her dedicated and determined coordination of the data collection logistics from beginning to end. We thank Alissa Shethar for assisting in all aspects of instrument development and data collection, as well as Carol Verboncoeur, Alyson Spencer-Reed, and Rebecca Brown for their help with the instruments, data management, and project administration. Throughout the study, we benefited from the expertise and technical assistance provided by Andrew Falk from UC Berkeley, who contributed to the observation and interview protocols and played an active role in collecting classroom data.

References

American Association for the Advancement of Science. (1993). Benchmarks for science literacy. New York: Oxford University Press. Retrieved from http://www.project2061.org/tools/benchol/bolframe.html. Accessed December 1, 2010.

Aschbacher, P., & Alonzo, A. (2006). Examining the utility of elementary science notebooks for formative assessment purposes. Educational Assessment, 11(3–4), 179–203.

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

Ball, D. L., Sleep, L., Boerst, T., & Bass, H. (2009). Combining the development of practice and the practice of development in teacher education. Elementary School Journal, 109, 458–474.

Barnett-Clarke, C., & Ramirez, A. (2004). Case discussions. In L. B. Easton (Ed.), Powerful designs for professional development (pp. 75–84). Oxford, OH: National Staff Development Council.

Birman, B., Desimone, L., Porter, A., & Garet, M. (2000). Designing professional development that works. Educational Leadership, 57(8), 28–33.

Black, P., Harrison, C., Lee, C., Marshall, B., & Williams, D. (2004). Working inside the black box: Assessments for learning in the classroom. Phi Delta Kappan, 86(1), 8–21.

Blank, R. K., de las Alas, N., & Smith, C. (2007). Analysis of the quality of professional development programs for mathematics and science teachers: Findings from a cross-state study. Washington, DC: Council of Chief State School Officers.

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

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

Boruch, R. F., DeMoya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (Eds.), Evidence matters: Randomized trials in education research (pp. 50–79). Washington, DC: Brookings Institution Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). How people learn. Brain, mind, experience, and school: Expanded edition. Washington, DC: National Academy Press.

Carlsen, W. S. (1993). Teacher knowledge and discourse control: Quantitative evidence from novice biology teachers' classrooms. Journal of Research in Science Teaching, 30, 471–481.

Carolina Curriculum for Science and Math. (2010). STC/MS: Science and technology concepts for middle schools. Burlington, NC: Author. Available at: http://www.stcms.si.edu.

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

Daehler, K. R., & Shinohara, M. (2001). A complete circuit is a complete circle: Exploring the potential of case materials and methods to develop teachers' content knowledge and pedagogical content knowledge of science. Research in Science Education, 31(2), 267–288.

Delta Education. (2010). FOSS: Full option science system (3rd ed.). Nashua, NH: Author. Available at: http://www.FOSSweb.com.

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

Dewey, J. (1965). The relation of theory to practice in education. In M. Borrowman (Ed.), Teacher education in America: A documentary history (pp. 140–171). New York: Teachers College Press. Original work published 1904.

Druva, C. A., & Anderson, R. D. (1983). Science teacher characteristics by teacher behavior and by student outcome: A meta-analysis of research. Journal of Research in Science Teaching, 20(5), 467–479.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: National Academies Press.

Engelhardt, P., & Beichner, R. (2004). Students' understanding of direct current resistive electrical circuits. American Journal of Physics, 72, 98–115.

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

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

Fulp, S. L. (2002). 2000 National Survey of science and mathematics education: Status of elementary school science teaching. Chapel Hill, NC: Horizon Research, Inc. Retrieved from http://2000survey.horizon-research.com/reports/elem_science.php.

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., & Bernbaum, D. (2006). Developing expertise with classroom assessment in K-12 science: Learning to interpret student work. Interim findings from a 2-year study. Educational Assessment 11(3 & 4), 237–263.

Gerard, L. F., Spitulnik, M., & Linn, M. C. (2010). Teacher use of evidence to customize inquiry science instruction. Journal of Research in Science Teaching, 47(9), 1037–1063.

Greenleaf, C., Hanson, T., Herman, J., Litman, C., Madden, S., Rosen, R., Kim-Boscardin, C., Schneider, S., & Silver, D., (2009). Integrating literacy and science instruction in high school biology: Impact on teacher practice, student engagement, and student achievement. Final report to the National Science Foundation. Grant #0440379.

Grossman, P. (2005). Research on pedagogical approaches. In M. Cochran-Smith & K. M. Zeichner (Eds.), Studying teacher education (pp. 425–476). Mahwah, NJ: Lawrence Erlbaum.

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. American Educational Research Journal 45(1), 184–205. DOI: 10.3102/0002831207312906. Retrieved from http://aer.sagepub.com/content/45/1/184.

Hashweh, M. (1987). Effects of subject matter knowledge in the teaching of biology and physics. Research and Teacher Education, 3, 109–120.

Heller, J. I., Daehler, K., & Shinohara, M. (2003). Connecting all the pieces: Using an evaluation mosaic to answer an impossible question. Journal of Staff Development, 24, 36–41.

Heller, J. I., Daehler, K. R., Shinohara, M., & Kaskowitz, S. R. (2004). Fostering pedagogical content knowledge about electric circuits through case-based professional development. Paper presented at the annual meeting of the National Association for Research on Science Teaching, Vancouver, B.C., Canada.

Heller, J. I., & Kaskowitz, S. R., (2004). Final evaluation report for Science Cases for Teacher Learning: Impact on teachers, classrooms, and students, Project years 2000–2003. Technical report submitted to Stuart Foundation.

Hewson, P. W., Kahle, J. B., Scantlebury, K., & Davis, D. (2001). Equitable science education in urban middle schools: Do reform efforts make a difference? Journal of Research in Science Teaching, 38(10), 1130–1144.

Hill, H., & Ball, D. L. (2009). The curious and crucial case of mathematical knowledge for teaching. Phi Delta Kappan 91(2), 68–71.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. American Educational Research Journal, 42(2), 371–406.

Jacob, R., Zhu, P., & Bloom, H. (2010). New empirical evidence for the design of group randomized trials in education. Journal of Research on Educational Effectiveness, 3(2), 157–198.

Kanter, D. E., & Konstantopoulos, S. (2010). The impact of a project-based science curriculum on minority student achievement, attitudes, and careers: The effects of teacher content and pedagogical content knowledge and inquiry-based practices. Science Education, 94, 855–887. DOI: 10.1002/sce.2039.

Kazemi, E., & Franke, M. L. (2004). Teacher learning in mathematics: Using student work to promote collective inquiry. Journal of Mathematics Teacher Education, 7(3), 203–235.

Lampert, M. (2010). Learning teaching in, from, and for practice: What do we mean? Journal of Teacher Education, 61(1–2), 21–34.

Lee, O. (2002). Science inquiry for elementary students from diverse backgrounds. In W. G. Secada (Ed.), Review of research in education. Vol. 26 (pp. 23–69). Washington, DC: American Educational Research Association.

Lee, O., & Fradd, S. H. (2001). Instructional congruence to promote science learning and literacy development for linguistically diverse students. In D. R. Lavoie & W-.M. Roth (Eds.), Models for science teacher preparation: Bridging the gap between research and practice (pp. 109–126). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Little, J. W. (2004). Looking at student work' in the United States: Countervailing impulses in professional development. In C. Day & J. Sachs (Eds.), International handbook on the continuing professional development of teachers (pp. 94–118). Buckingham, UK: Open University Press.

Little, J. W., Gearhart, M., Curry, M., & Kafka, J. (2003). Looking at student work for teacher learning, teacher community and school reform. Phi Delta Kappan, 85(3), 184–192.

Little, J. W., & Horn, I. S. (2007). ''Normalizing'' problems of practice: Converting routine conversation into a resource for learning in professional communities. In L. Stoll & K. S. Louis (Eds.), Professional learning communities: Divergence, depth and dilemmas (pp. 79–92). London, England: Open University Press.

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

McDonald, J. P. (2001). Students' work and teachers' learning. In A. Lieberman & L. Miller (Eds.), Teachers caught in the action: Professional development that matters (pp. 209–235). New York: Teachers College Press.

Mundry, S., & Stiles, K. E. (Eds.). (2009). Professional learning communities for science teaching. Arlington, VA: NSTA Press.

National Research Council. (2002). Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools. Washington, DC: National Academy Press.

National Research Council, National Committee for Science Education Standards and Assessment. (1996). National science education standards. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog/4962.html.

Partnership for 21st Century Skills. (2008). NSTA 21st century skills, education and competitiveness: A resource and policy guide. Washington, DC: Author.

Quellmalz, E., Timms, M., & Buckley, B. (2005). Using science simulations to support powerful formative assessments of complex science learning. San Francisco, CA: WestEd.

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. (2011). Videobased lesson analysis: Effective science PD for teacher and student learning. Journal of Research in Science Teaching, 48(2), 117–148. DOI: 10.1002/tea.20408.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. Journal of Research in Science Teaching, 44(1), 57–84.

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

Scher, L., & O'Reilly, F. (2009). Professional development for K-12 math and science teachers: What do we really know? Journal of Research on Educational Effectiveness, 2(3), 209–249.

Schneider, R. M., & Plasman, K. (2011). Science teacher learning progressions: A review of science teachers' pedagogical content knowledge development. Review of Educational Research, 81(4), 530–565. DOI: 10.3102/0034654311423382.

Shinohara, M., Daehler, K. R., & Heller, J. I. (2004). Using a pedagogical content framework to determine the content of case-based teacher professional development in science. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Vancouver, B.C., Canada.

Shipstone, D. (1988). Pupils' understanding of simple electrical circuits. Some implications for instruction. Physics Education, 23(2), 92–96.

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

Shulman, L. S. (2005). Teacher education does not exist. Stanford Educator, p. 7.

Shymansky, J., & Matthews, C. (1993). Focus on children's ideas about science: An integrated program of instructional planning and teacher enhancement from the constructivist perspective. The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics. Ithaca, NY: Misconceptions Trust.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. Educational Researcher, 31(7), 15–21.

Sloan, H. A. (1993). Direct instruction in fourth and fifth grade classrooms. Dissertation Abstracts International, 54(8), 2837A. UMI No. 9334424.

Tosun, T. (2000). The beliefs of preservice elementary teachers toward science and science teaching. School Science and Mathematics, 100, 374–379. DOI: 10.1111/j.1949-8594.2000.tb18179.x.

Van Driel, J. H., Verloop, N., & De Vos, W. (1998). Developing science teachers' pedagogical content knowledge. Journal of Research in Science Teaching, 35(6), 673–695.

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

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education (Research Monograph No. 6). Council of Chief State School Officers and National Institute for Science Education. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research. Retrieved from http://hub.mspnet.org/index.cfm/9874.

Webb, N. L., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2006). The WEB alignment tool: Development, refinement, and dissemination. In Council of Chief State School Officers (Ed.), Aligning assessment to guide the learning of all students: Six reports (pp. 1–30). Washington, DC: Author.

Weiss, I. R., & Miller, B. (2006). Developing strategic leadership for district-wide improvement of mathematics education. Lakewood, CO: National Council of Supervisors of Mathematics.

White, B., Frederiksen, J., & Collins, A. (2009). The interplay of scientific inquiry and metacognition: More than a marriage of convenience. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), Handbook of metacognition in education (pp. 175–205). Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, S. M., Rozelle, J. J., & Mikeska, J. N. (2010). Cacophony or embarrassment of riches: Building a system of support for quality teaching. Journal of Teacher Education 62(4), 383–394. DOI: 10.1177/0022487111409416.

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/regions/southwest/pdf/REL_2007033.pdf.