Investigating How and Under What Conditions Effective Professional Development Increases Achievement in Elementary Science

Executive Summary of the Final Report to the Institute of Education Sciences[1]

Judith Warren Little, Elena Durán López, Anna Weltman
University of California, Berkeley

Joan I. Heller, Nicole Wong, Selena Burns, and James Owen Limbach
Heller Research Associates

Luke Miratrix and Lo-Hua Yuan
Harvard University

Contact: Nicole Wong, nwong@gordonheller.com

The goal of this Institute of Education Sciences (IES)-funded project was to further the field's understanding about precisely how professional development (PD) for teachers can result in changes in student learning. Through a fine-grained analysis of a large corpus of video, interview, and survey data from both PD and classroom instruction, we characterized teachers' own learning experiences and their instructional practices when teaching 4th grade science. This study generated hypotheses about relationships among teachers' participation in science PD, changes in their content knowledge and pedagogical content knowledge (PCK), their classroom practice, and their students' learning.

The study included secondary analyses of qualitative and quantitative data generated from the Learning Science for Teaching (LSFT) project[2], a randomized controlled trial comparing three variations of a PD course focused on grade 4 teaching of electric circuits. The eight national sites in the study included more than 280 elementary teachers and nearly 7,000 students. In three "intensive" sites, the research team invested additional resources in videotaping PD activity and classroom instruction of

[1] This document summarizes findings for the *Investigating How and Under What Conditions Effective Professional Development Increases Achievement in Elementary Science* final report submitted to the Institution of Education Sciences for grant #R305A150341. To obtain the full report and related publications, contact nwong@gordonheller.com.

[2] *Effects of Content-Focused and Practice-Based Professional Development Models on Teacher Knowledge, Classroom Practice and Student Learning in Science,* Teacher Professional Continuum Program, National Science Foundation Grant No. 0545445.

randomly selected focal teachers in order to support a qualitative analysis of the relationship between participation in the PD and classroom practice. (For a comprehensive account of the study's research design, data sources, and data collection methods, see Heller, Little, & Shinohara, 2010.)

The LSFT project compared three variations of the *Making Sense of SCIENCE (MSS): Electric Circuits* course, one of many courses in the WestEd teacher professional development series. The course variations engaged teachers in identical Science Investigations designed to support teachers' content knowledge, but they differed in the supports for teachers' PCK as indicated by the following treatment designations: A) Teaching Cases; B) Looking at Student Work; and C) Metacognitive Analysis[3]. Control teachers participated in "business as usual," attending any school-provided or district-provided PD that may have been available during the study period. As a requirement for participation in the study, all teachers committed to teaching a unit[4] on electric circuits[5] to their 4th grade students using their district-provided curricula, such as *Full Option Science System* (FOSS), *Science and Technology Concepts* (STC), or other teacher-sourced or teacher-generated materials. The MSS courses did not include an accompanying student curriculum, but teachers in Treatment B (Looking at Student Work) were given access to samples of student tasks that could be used in their classrooms to elicit artifacts of student thinking for collaborative analysis during their PD course.

Previous analyses from the LSFT study demonstrated statistically significant and lasting gains in teacher knowledge and student learning for all three experimental conditions relative to controls (Heller, Little, & Shinohara, 2010; Heller et al., 2012). In the current work, we capitalized on LSFT's comprehensive qualitative and quantitative data set to investigate *why* and *how* the PD worked to produce these positive results. In doing so, we sought to improve the explanatory power of a conceptual model of PD effects and to inform efforts to scale up promising PD models.

3 Although course materials for treatment C were titled *Content Immersion*, we use *Metacognitive Analysis* to highlight the course emphasis on teachers' analysis of their own learning.

4 In a published review of elementary science programs, Slavin et al. (2014) argued that aspects of the research design in this study, as reported by Heller et al., 2012, made a comparison of treatment and control groups inappropriate. Slavin et al.'s criticism would have been justified had their portrayal of the research design been accurate, but it was not. We therefore emphasize that all teachers—treatment and control—taught an electric circuits unit in their 4th grade classrooms and all used the local curriculum available to them. The *Making Sense of SCIENCE: Electric Circuits* PD constitutes a curriculum for adults, not one designed for elementary students.

5 At the time of the study, electric circuits topics were included in state standards for 4th grade at all of the study sites.

Research Questions

The basic theory of action underlying all MSS courses (*Figure 1*) posits that MSS teacher PD interventions improve teacher knowledge, which changes classroom practice, which in turn improves student achievement. This study addressed three sets of research questions that form a theoretical bridge among these key components, to begin unpacking why and how these changes occur.

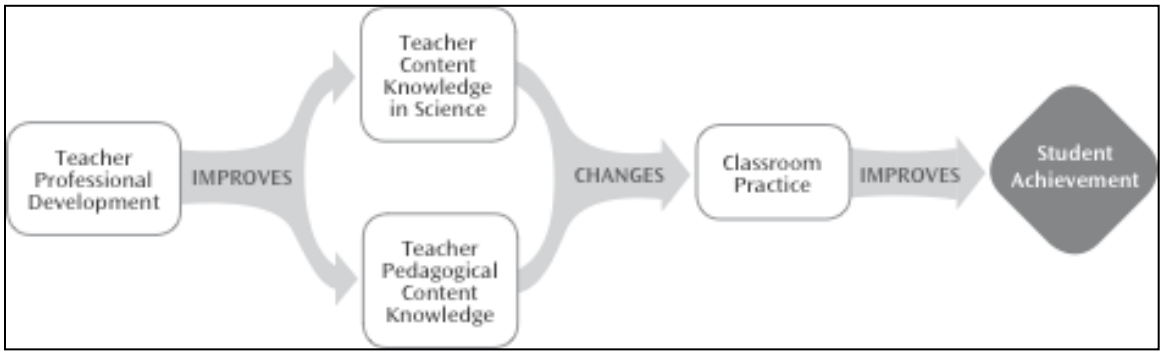

Figure 1. Making Sense of SCIENCE professional development theory of action. Adapted from Horizon Research's ATLAST Theory of Action http://www.horizon-research.com/atlast

The goal of the first set of questions was to help us understand the teachers' opportunity to learn during the PD, as enacted. These questions focused on variations in teacher interaction and facilitation practice in the PD courses within and across the three treatment conditions.

1. What patterns of teacher interaction, with respect to teachers' collective engagement in scientific sense-making and selected aspects of teaching practice, characterized the PD courses?
2. What features of facilitation practice were most associated with teachers' observed patterns of interaction? Specifically, how and to what extent did facilitators (a) focus on conceptual learning goals; (b) help build accuracy, breadth, and depth of science content understanding by, for example, incorporating visual representations of data and prompts for small and whole-group discussions of evidence; and (c) support the development of content-related teaching knowledge by, for example, guiding a focus on student thinking or by inviting a critical analysis of particular instructional practices?
    1.1. How did newly trained and expert facilitators differ in implementing the PD?
    1.2. Are there differences in the quality of PD facilitation between Round 1 and Round 2 of implementation?
3. What features of facilitator training and facilitator background may have influenced facilitation practices and other features of the enacted PD?

A second set of questions explored the relationship between the PD implementation, teacher knowledge, classroom practice, and student achievement.

First, we sought to determine whether and how treatment and control teachers' classroom practices reflected a robust conception of science teaching and learning by asking:

4. What teaching practices and patterns of student interaction characterize each classroom lesson observed, with respect to (a) the level of student cognitive engagement in science, (b) classroom focus on conceptual understanding of core science ideas, (c) use of visual representations and models to support learning, (d) student engagement in scientific sense-making practices, such as evidence-based reasoning, and (e) teachers' elicitation of and attention to student thinking. How did treatment and control teachers' lessons compare with respect to classroom practices and patterns of student interaction?

Next, we explored various relationships among the qualitative PD ratings, teacher content knowledge, classroom ratings, and student content knowledge through quantitative analysis and qualitative case studies.

5. What relationships exist between patterns of classroom interaction and student content knowledge?
6. What relationships exist between features of the PD and classroom patterns of interaction? What aspects of teachers' experiences in the PD may be most effective in strengthening their content-related teaching knowledge and classroom practices?
7. What features of PD implementations are associated with stronger impacts on (a) teachers' content knowledge and (b) students' content knowledge? To what extent do these PD features account for these impacts?

The third set of questions focused on the identification and statistical testing of moderator and mediator variables that may have strengthened or weakened the impact of the PD on teacher and student outcomes. We asked:

8. Which teachers, in which school and district contexts, benefited the most from the PD? More specifically, which subgroups defined by teachers' science content knowledge, science background, or teaching experience, benefited most? How did school variables such as local opportunities for collegial interaction among science teachers, or district resources for science education, influence impact of the PD?
9. Under what conditions were PD impacts on student achievement the strongest? More specifically, which subgroups defined by student demographics, teachers' pre- or post-PD science content knowledge, or teachers' science background or teaching background, benefited most? How did conditions, defined by school and district context variables, influence impact of the PD?
10. Did the PD improve student outcomes by producing effects on teacher knowledge and skill? What teacher outcomes mediated impacts of the PD on student gain scores?

Data Sources

Data sources for the project mapped against the presumed causal chain of influences from PD to student outcomes (see *Figure 1*) and supplied contextual information about teachers, facilitators, and students.

Data on the PD interventions came from 24 courses that took place over two rounds at eight national sites. At one site, PD was led by expert facilitators from WestEd who developed the courses. PD at the remaining seven sites was led by newly trained local facilitators. Data included video recordings of facilitator training; facilitator background surveys; one-time tests of facilitator content knowledge, conducted after facilitator training but prior to implementation; video recordings of PD sessions from all sites; facilitator interviews; debriefings with project staff throughout PD implementation; and post-implementation facilitator focus groups.

Data for the study of classroom practices came from 30 randomly selected focal teachers from the experimental and control groups. Data included video recordings of two consecutive lessons about electric circuits; pre-and post-observation interviews about the reasoning underlying teachers' planning and teaching practices; pedagogical content knowledge interviews administered at three points during the study; video of the focal teachers' participation in PD; classroom context information from teacher background surveys, post-PD, and post-instruction surveys; teacher pre- and post-PD content knowledge tests; and student pre- and post-instruction content knowledge tests.

Data for the analysis of moderating and mediating variables came from the entire corpus of 280 elementary teachers and nearly 7,000 students in the study. Data included pre-PD teacher background surveys, teacher post-PD surveys, teacher post-instruction surveys, teacher pre- and post-PD content knowledge tests; and student pre- and post-instruction content knowledge tests.

Analysis of PD features and variation in course implementation

Variations in PD course quality. We generated two summary PD quality ratings for 12 of the 24 courses, four led by expert (WestEd) facilitators and eight led by local, newly trained facilitators. The rated courses represented all eight national research sites, and four courses each from Treatments A (Teaching Cases), B (LASW), and C (Metacognitive Analysis). We focused principally on the quality of facilitation and the nature of teacher interaction in whole-group activity in the first three of each course's eight sessions. These sessions approximated the content likely to be taught in fourth grade classrooms and thus provided a likely parallel to the analysis of classroom instruction. By the third session, the groups also had sufficient time to develop a way of working together, enabling us to determine whether there was a pattern of collective sense-making.

Rating 1 assessed the quality of support for teachers' science content learning, with ratings focused on fidelity to designed activities; extent of teacher participation; and supports for content depth through teachers' collaborative science sense-making, the

use of visual representations, and the cumulative development of a visible public record of discovery. Rating 2 assessed the quality of support for developing science knowledge for teaching (PCK), with ratings focused on the instructional approach framed and modeled by facilitators; the nature and extent of teachers' focus on aspects of student thinking (relevant only to Treatments A and B); and the explicit attention given to instructional decision making and practice (all treatments).

Rating 1 resulted in four courses clustered at the low end (ratings 1.5–2.0); four occupying a middle range (3.0–3.5), and four clustered at the top (4.5–5.0). Courses at the top end of the range were all facilitated by the expert facilitators. However, a rating of 3.0 and above reflected a solid level of facilitation quality. At the high end of the scale, facilitators employed multiple moves to engage teachers in careful consideration of evidence from small-group activity, encouraged them to develop general claims anchored to the evidence, and made time for teachers to identify and work on points of uncertainty or confusion. At the low end of the scale, facilitators did little to engage teachers themselves in collaborative sense-making during whole-group activity.

Rating 2 produced profiles differentiated by treatment. Of the four Treatment C courses, two were ranked low (1.5), one was ranked in the middle (3.0), and one was ranked moderately high (4.0). Of the Treatment A and Treatment B courses, those focusing centrally on attention to student thinking and the analysis of student work, one course was rated low (2.0) and all others received ratings at mid-range (3.0–3.5) or above, with five courses clustered at or near the top (4.5–5.0).

Of the eight courses judged to be of moderate to high quality on both ratings, four were led by WestEd expert facilitators and four were led by local site facilitators who were recruited, trained, and supported by WestEd. Courses with moderate-to-high ratings maintained a conceptual focus, engaged teachers in collaborative scientific reasoning, and made good use of visual representations to support learning. PCK supports were especially evident in the two treatments that focused on analyzing details of student work and considering implications for classroom instruction. Overall, however, supports for teachers' own science content learning were found to be stronger than the supports for other aspects of their science teaching (such as the tradeoffs among instructional choices or anticipating likely student responses).

Comparing courses led by WestEd expert facilitators and newly trained local facilitators. A comparison of courses led by WestEd expert facilitators and those led by newly trained local facilitators identified four aspects of expert facilitation that were tightly integrated in practice. Expert facilitators: (1) consistently anchored whole-group discussion in publicly displayed data from small-group science investigation; (2) employed a range of visual representations and inscriptions to support teachers' scientific sense-making and to structure participation; (3) created a dynamic of sustained and widespread teacher participation in whole-group discussions, enabling teachers to do the scientific sense-making that culminated in an understanding of key concepts; and (4) worked to orient discussions of science teaching to the shared PD materials, examples, and experiences appropriate to each treatment, ensuring depth and specificity. Newly trained local facilitators in higher rated courses, while not

matching the sophistication of the expert facilitators, approximated the experts' practices with regard to a sustained conceptual focus, consistent efforts to engage teachers in doing the science sense-making, the use of multiple visual representations to support that sense-making and teacher learning, and (in Treatments A and B) the approach taken to discussions of student thinking and instruction. Facilitators of lower rated courses varied with respect to the challenges they exhibited, but those challenges included a loss of conceptual focus, limited attention to evidence, a tendency to provide information or explanation rather than invite teachers' own sense-making, and a somewhat fragmented and under-developed use of visual representations to support learning. Facilitation practice throughout was consistent with the PD design structure, but not always consistent with the stance of facilitation practice modeled and promoted in the facilitator training and in the Facilitation Guide. Results of the PD course analysis were used to generate a set of hypotheses about the potential influence of the PD on the classroom, which are described in the full report.

Comparing Round 1 and Round 2 courses. In the main effects analysis reported previously, no statistical differences were found to differentiate Round 1 and Round 2 results for teachers and students (Heller et al., 2010). We were interested in whether the qualitative data showed differences in PD quality, but given limited time and resources, we confined this analysis to two facilitator pairs and their respective Round 1 and Round 2 courses. We found no change in the Round 2 facilitation for a facilitator pair whose course was lower rated in Round 1. However, facilitators whose course had been rated in the medium-to-high range in Round 1 showed indications that their detailed attention to teachers' thinking in the Round 1 course influenced the approach they took to some key concepts in Round 2.

The contribution of facilitator background, facilitator training, and other resources to PD quality. In analyzing the facilitator training video, we distinguished between direct (explicit) and indirect (implicit) preparation for facilitation. Direct preparation refers to activities in which the participants were invited to take on "the facilitator's hat" (for example, discussion of sample course facilitation video, or guided segment planning and rehearsal). Indirect preparation refers to the facilitation practices that the expert WestEd facilitators modeled as they conducted the Science Investigation and the treatment-specific PCK segments. Overall, the facilitator training relied more heavily on the indirect preparation achieved by modeling an approach to PD that emphasized collaborative inquiry and sense-making. The WestEd facilitators also introduced moments of explicit framing (articulation of principle), facilitation advice, or explication of facilitation moves. Those embedded moments of explication addressed the key aspects of the theory of action that underlies the PD design and that formed the basis of the PD rating scales: (1) fostering teacher participation and supporting teachers themselves in learning from and with each other; (2) achieving an in-depth exploration of key science ideas and engaging in rich scientific sense-making; and (3) developing new resources and practices for science teaching. The new facilitators rated the training highly and credited it for both extending their content knowledge and building their confidence for implementing the PD. There is some indication that facilitators' background in leading PD may have contributed to the differentiation between the medium- and low-rated courses; the facilitators of the medium-rated

courses had all received prior preparation specifically focused on the work of facilitation.

Analysis of impact of PD on classroom instruction

We rated 30 classroom videos on a scale of 1–5 for each of the five dimensions of the Heller Research Associates (HRA) Classroom Rating Scale: (1) students are cognitively engaged in science; (2) the classroom experience is focused on conceptual understanding of core science ideas; (3) representations suited to science sense-making are used to support learning; (4) students are engaged in scientific sense-making practices; and (5) the teacher elicits and attends to student thinking. We believe that classrooms that earn high ratings on these five dimensions are ones that promote student-centered, collaborative, conceptual sense-making in science.

The Harvard team conducted a principal component analysis and found that the first component (PC1) accounted for 70% of the variability in the data. This weighted average of the five dimensional scores was used as a proxy for "overall classroom quality." Sensitivity checks permuting a simple average of the five scores and the five dimensional ratings corroborated the significant findings based on PC1.

The PD had a large, statistically significant effect on the overall quality of classroom instruction. Permutation tests controlling for location and teachers' pre-PD knowledge of electric circuits were run to test whether there were significant differences in overall classroom quality between control and treatment classrooms. When we pooled the scores from each of the treatment variations (PD Treatments A, B, and C) and compared them to the control condition, we detected large, statistically significant ratings of classroom lessons' overall quality (Est. TrtEffect = 2.390, $p = .004$).

When compared to control, ratings for quality of classroom instruction were higher for each course variation, but these PD effects reached statistical significance for only two of the three courses: Looking at Student Work (B) and Metacognitive Analysis (C). When controlling for location and teachers' pre-PD content knowledge, permutation tests showed large, statistically significant differences in overall classroom quality between Treatment B vs control (Est. TrtEffect = 2.305, $p = .020$) and Treatment C vs control (Est. TrtEffect = 2.440, $p = .014$). Differences between Treatment A vs control did not reach significance ($p > 0.05$), possibly due to an outlier.

Teacher participation in the PD resulted in statistically-significant improvements in student cognitive engagement, student engagement in scientific sense-making practices, and teacher elicitation of and attention to student thinking. Permutation tests were conducted to determine whether there were significant differences between control and treatment classrooms for each of the five analytic dimensions in the HRA Classroom Scoring Rubric. After controlling for location and teachers' pre-PD electric circuits knowledge and adjusting for multiple comparisons, where the p-value threshold is .008, significant differences were found for three of the dimensions (Est. TrtEffect = 0.992, $p = .004$ for cognitive engagement; Est. TrtEffect = 0.956, $p = .008$ for science practices; and Est. TrtEffect = 0.998, $p = .006$ for elicitation of and attention to student thinking). Large differences were observed between the treatment and control

groups for classroom focus on conceptual understanding and use of representations, but these differences were not statistically significant at the .008 level. (Est. TrtEffect = 1.025, p = .048 for conceptual understanding, and Est. TrtEffect = 1.071, p = .024 for representations).

Teachers participating in the same treatment condition varied in their classroom practice. The distribution of individual classroom ratings showed some degree of variation among teachers participating in the same treatment. For example, the average classroom rating for Treatment A teachers in the Site 6 courses clustered into two groups: one group with lower scores (between 3 and 4) and another with higher scores (between 4 and 5). Although these teachers experienced the same treatment, they did not all teach in the same way. Informed by these results and guided by a set of hypotheses derived from the analysis of the PD courses about what elements of the PD could be traced to classroom practice, we conducted a set of case studies to investigate both the influence of the PD on classroom practice within the five dimensions, and variation among teachers who had participated in the same PD. These case studies confirmed the impact of the PD on teacher practice across all dimensions, including revealing ways in which specific pedagogical practices intended to elicit conceptually based scientific sense-making were traceable from the PD to treatment teacher classrooms. However, the case studies also showed that the degree of this impact varied, based in part by how closely teachers followed their district-provided curricula.

Analysis of the relationship between PD quality, teacher knowledge, classroom ratings, and student knowledge gains

Ratings of overall classroom quality were positively correlated with student content knowledge gains and teacher content knowledge. Beyond what can be explained by teachers' pre-PD content knowledge, higher classroom ratings were significantly correlated with higher student gains. When controlling for teachers' pre-PD content knowledge, we found a moderate relationship between overall classroom quality and gains on student content knowledge test scores (r = .37, p = .06). We also found a correlation between overall classroom quality and both teacher quiz 2 (r = .39, p = .03) and teacher quiz gain (r = .47, p = .01) when teacher quiz 1 is controlled for, suggesting a relationship between teacher's post-PD knowledge, teacher learning, and overall classroom quality. These results are consistent with the hypothesis that better teacher content knowledge is correlated with teaching practices that create a classroom environment conducive to scientific sense-making. They also suggest that our rubric is, indeed, measuring classroom practices that are associated with gain in student content knowledge.

PD course quality was positively correlated with student content knowledge gains. We found a statistically significant positive relationship between the PD quality indicated by the PD video ratings and student learning outcomes, as measured by gains on the student test. More specifically, we found a significant correlation between Rating 1 (support for teachers' content learning) and student learning outcomes, but not for

Rating 2 (support for the teaching of content, or PCK). These findings are robust to a variety of different model specifications.

PD course quality was not correlated with teacher content knowledge gains or classroom quality ratings. Given the relationships at the student level, we surprisingly did not find significant association between either of the ratings of PD quality and gains in teachers' own content knowledge. These null findings are again consistent across a variety of modeling specifications. Additionally, higher PD ratings were not correlated with higher classroom quality ratings after controlling for teachers' pre-PD content knowledge.

Analysis of moderating and mediating variables

Numerous candidate moderators and mediators of impact on student and teacher outcomes were explored, including teacher experience and PD background in teaching science and electric circuits in particular; classroom and school context (e.g., percent of students who were English language learners or qualified for free or reduced-price meals in the classroom, and support for science teaching, including opportunities for collaborating with other teachers around science teaching), and student demographic variables (race/ethnicity, gender, and English language proficiency). The full list of variables and the instruments that served as sources of each are listed in the full report, along with a full account of the analysis methods. For all moderation analyses, we pooled data from participants in PD Treatments A, B, and C to comprise the treated group, while group D was control.

Moderation of teacher outcomes. Two covariates were found to be moderators of teacher test outcomes (significant at $p \leq .05$): the number of years the teacher taught science (among treated teachers, teacher gain was larger for teachers who had taught fewer years of science), and the teacher content pretest scores (among treated teachers, impact on teacher posttest scores was larger for teachers whose pretest scores were lower). Because these were exploratory analyses, we also note that two other variables showed marginally significant results at $p \leq .10$: whether the teacher had ever taught an electric circuits unit before the study year (among treated teachers, teacher gain was larger for teachers who had never previously taught electric circuits), and whether the teacher reported school or district support for science teaching (gains were greater for teachers who did not report support for science teaching). Teachers with the least experience, knowledge, and prior support for science teaching benefited the most from the PD.

Moderation of student outcomes. We found one covariate to be a *moderator of student gains*: whether their teacher taught an electric circuits unit in the most recent year preceding the study (among treated teachers, student gain was larger for teachers who had *not* recently taught an EC unit compared to teachers who did teach EC prior to the study). Again, because these were exploratory analyses, we also note that two other variables showed marginally significant results at $p \leq .10$: the number of hours of science teaching PD their teachers received in the last three years (in classes of treated teachers, student gain was greater for teachers with fewer hours of previous science

PD), and the proportion of students in the teacher's class that were English learners (in classes of treated teachers, student gain was larger in classrooms with higher proportions of English learners). Overall, students who were most disadvantaged benefited the most from their teachers receiving the PD.

Mediation of student outcomes. There was strong evidence that the PD increased teachers' scores on written assessments of both science content knowledge and a written assessment of PCK. The impact of the PD on teacher content knowledge was found to partially mediate the effect of PD on student gain scores, indicating that some, but not all, of the impact of PD on student gain may be explained by teachers' science content knowledge. Overall, the PD's great impact on teachers' relevant science knowledge did benefit their students, but other factors likely also mediated the PD influence.

There was strong evidence that teachers' post-PD PCK scores were predictive of student gains, which suggests that PCK plays a role in improving student outcomes. However, the impact of the PD on PCK (the change in PCK) was not found to be a significant mediator of student outcomes. The non-significant mediator analysis may indicate that the written PCK measure we used was not sensitive to the full impact of the PD on PCK.

Implications

This study explored a rich and comprehensive corpus of video, interview, and survey data in an effort to specify how and under what conditions a demonstrably successful program of science PD achieved its positive outcomes for teacher and student learning. In doing so, we also sought to "connect the dots" in a conceptual model of PD that envisions a cascade of influences from features of the PD to direct impact on teacher knowledge, intermediate impact on classroom learning environments, and finally the more distal effects on student achievement (Cohen & Hill, 2000; Desimone, 2009; Heller, Daehler, & Shinohara, 2003; Scher & O'Reilly, 2009; Weiss & Miller, 2006).

An in-depth qualitative examination of the PD implementation enabled us to see how the conceptual orientation and collaborative sense-making practices were instantiated in each course. Altogether, the analysis supplied evidence of a PD experience broadly consistent with the intended design (despite variations) and helped to explain how it was that the treatment teachers and their students significantly outperformed controls.

We found strong evidence that the PD had a large, positive impact on classroom practice. Furthermore, we found evidence of a moderate, significant correlation between classroom practice and student outcomes. In a previous analysis of data from this study (Heller et al., 2012), we found that treated teachers and their students had greater content knowledge gains than the control group, but we had not yet characterized the PD conditions that led to teacher learning or the classroom learning conditions that may have led to improvements in student learning. The findings in the

current analysis provide support for our hypothesized theory of action by demonstrating that teachers' participation in MSS PD resulted in higher quality science instruction along dimensions that were correlated with better student outcomes.

This study stands as one of a small but growing number of studies that employ an experimental design with random assignment to study the impact of teacher PD on teacher knowledge, teaching practice, and student learning (National Academies of Science, 2015). In addition, the study is distinctive in two significant respects: (a) its scale, involving a large number of teachers and students in multiple sites; and (b) its ability, given the constellation of qualitative and quantitative data, to investigate the presumed causal relationships among teachers' experiences in PD, gains in teacher knowledge, changes in teacher practice, and student learning.

Our study provides evidence of the impact of PD on teachers' classroom practice and suggests how teachers' participation may have resulted in student learning. It also conveys some of the complexity of the presumed "cascade of influences" by which PD yields teacher and student outcomes. We maintain that these findings present an additional, empirically based challenge to the declared consensus regarding features of effective PD.

Desimone (2009) argued that research on teacher PD would be strengthened by systematic attention to selected design features that had been determined to be characteristic of effective PD, among them: content focus; provisions for "active participation;" and collective participation. Despite some criticisms of the empirical basis for these design features (Wilson, 2013) or their utility for investigating PD (Kennedy, 2016), they have remained a prominent point of departure for research on PD. Yet most studies treat those design features as proxies for the quality of teacher experience. With rare exception (e.g., Grigg et al., 2013), studies of PD do not delve into the enacted pedagogy of PD and the learning opportunities constructed for and by teachers, nor do they trace the central emphases of the PD into the classroom. Our findings demonstrate the insufficiency of design features as a proxy for PD quality or to account for outcomes; rather, our findings underscore the importance of investigating the ways that PD design features unfold in practice, and to explore how variations in enactment relate to outcomes, including teachers' classroom practice.

Contributors

To conduct this work, Judith Warren Little (PI) at the University of California, Berkeley, collaborated with Joan I. Heller (co-PI) and Nicole Wong of Heller Research Associates and with Professor Luke Miratrix (co-PI) of Harvard University. Elena Durán López (UCB), Anna Weltman (UCB), Selena Burns (HRA), J. Owen Limbach (HRA), and Lo-Hua Yuan (Harvard) were all members of the research team and made significant contributions to the analysis and writing. We convened an Advisory Group to help us refine the research plan, determine analytic priorities, and weigh various sampling plans and methods of analysis. Members included: Hilda Borko (Professor, Stanford University), Anne M. Chamberlain (Senior Research Associate, Manhattan Strategy

Group), Corinne Herlihy (Project Director, National Center on Teacher Effectiveness, Harvard University), Elham Kazemi (Professor, University of Washington), and Catherine Lewis (Professor, Mills College). Mayumi Shinohara (Graduate Student, Vanderbilt University) and Kirsten R. Daehler (Director of Making Sense of SCIENCE, WestEd), who had designed the PD interventions and had been key participants in the prior NSF study, also served on the Advisory Group and offered additional consultation throughout the project.

References

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

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

Grigg, J., Kelly, K. A., Gamoran, A., & Borman, G. D. (2013). Effects of two scientific inquiry professional development interventions on teaching practice. *Educational Evaluation and Policy Analysis, 38*(1), 38-56.

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., Wong, N., Shinohara, M., & Miratrix, L. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching, 49*(3), 333– 362.

Heller, J. I., Little, J. W., & Shinohara, M. (2010). *Impact of content-focused and practice-based professional development models on elementary electric circuits teaching and learning.* Final Report to the National Science Foundation, Grant No. 0545445.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research, 86*(4), 945–980. doi:10.3102/0034654315626800

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.

Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*, 51(7), 870-901.

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

Wilson, S. (2013). Professional development for science teachers. *Science, 340*, 310-313.