NCEE 2012-4002

U. S. DEPARTMENT OF EDUCATION

Effects of Making Sense of SCIENCE™ professional development on the achievement of middle school students, including English language learners

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Effects of Making Sense of SCIENCE™ professional development on the achievement of middle school students, including English language learners

Final Report

March 2012

Author:
Joan I. Heller
Heller Research Associates

Project Officer:
OK-Choon Park
Institute of Education Sciences

NCEE 2012–4002
U.S. Department of Education

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
Rebecca A. Maynard
Commissioner

March 2012

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract ED-06C0-0014 with Regional Educational Laboratory West administered by WestEd.

Disclosure of potential conflicts of interest

Regional Educational Laboratory West, housed at WestEd, contracted with Heller Research Associates to conduct a third-party evaluation of Making Sense of SCIENCE™, a WestEd intervention. The author, other staff from Heller Research Associates involved in the study, and members of the Technical Work Group for the study have no financial interests that could be affected by the content of this report. The evaluation was conducted independent of WestEd staff, who developed and implemented Making Sense of SCIENCE™. [1]

[1] Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

Contents

Tables

Figures

Acknowledgments

The Regional Educational Laboratory (REL) West research team would like to acknowledge colleagues who made the study possible from the early design phases to the final analyses.

We thank the site coordinators who made the program implementation and teacher data collection possible: Peter A'Hearn, Karen Cerwin, Bree Watson, Kirstin A. Bittel, Joan Gilbert, Kathleen Blair, Paul Gardner, Dale Moore, Melissa Smith, and Nicole Wickler. We also thank the course facilitators, who so skillfully delivered all of the professional development courses: Peter A'Hearn, Kirstin A. Bittel, Meg Gebert, Kathleen Blair, Dan Lavine, Sarai Costley, Homeyra Sadaghiani, Sylvia Gutman, James Hetrick, Teresa Vail, and John Lazarcik.

We are very grateful to all of the teachers and students who contributed to this study. We recognize the burden associated with participating in a research study of this magnitude and thank them for their time, commitment, diligence, and interest over the past several years.

Colleagues at WestEd, the developers of Making Sense of SCIENCE™ professional development courses, worked with the research team for several years as the study design was developed and the intervention was provided to teachers. We acknowledge the unwavering commitment of the implementation team and all of the staff who supported the project: Mayumi Shinohara, Kirsten Daehler, Mikiya Matsuda, and Jennifer Mendenhall. We give a huge thank you to Cara Peterman at Heller Research Associates for her dedicated and diligent coordination of the data collection logistics from beginning to end. We also thank Alyson Spencer-Reed and Carol Verboncoeur for their help with the instruments, data management, and project administration.

Finally, the REL West team thanks the technical working group that provided guidance from the outset through to the final analyses: Jamal Abedi, University of California, Davis; Lloyd Bond, Carnegie Foundation for the Advancement of Teaching; Geoffrey Borman, University of Wisconsin; Brian Flay, Oregon State University; Tom Good, University of Arizona; Corinne Herlihy, Harvard University; Joan Herman, National Center for Research on Education, Standards, and Student Testing, University of California, Los Angeles; Heather Hill, Harvard University; Roger Levine, American Institutes for Research; Juliet Shaffer, University of California, Berkeley; and Jason Snipes, IMPAQ International.

Executive summary

This study evaluated an approach to professional development for middle school science teachers by closely examining one grade 8 course that embodies that approach. Using a cluster-randomized experimental design, the study tested the effectiveness of the Making Sense of SCIENCE™ professional development course on force and motion (Daehler, Shinohara, and Folsom 2011) by comparing outcomes for students of teachers who took the course with outcomes for students of control group of teachers who received only the typical professional development offered in their schools and districts. The study estimated impacts on student science achievement for all grade 8 students in the study sample as well as for the subsample of English language learners. It also estimated impacts on teacher science and pedagogical knowledge.

Need for better preparation of science teachers

Teacher courses developed by the Understanding Science for Teaching program at WestEd are, according to the developer, intended to improve students' science achievement, including that of low-performing students and English language learners, by strengthening their teachers' science content knowledge and knowledge for teaching that science. Making Sense of SCIENCE™ courses have been shown to increase elementary school teachers' content knowledge and student achievement in a national randomized experimental controlled trial and numerous smaller field tests (Heller, Daehler, and Shinohara 2003; Heller et al. 2010).

The need for better preparation of science teachers is clear: More than two-thirds of middle school science teachers in the United States reportedly have inadequate science preparation (Fulp 2002). "Out-of-field" teaching is widespread and stands to increase as many veteran science teachers retire (Fulp 2002). For example, one study reported that only 28 percent of science teachers in grades 6–8 have an undergraduate degree in science (Fulp 2002). Quality professional development for middle school teachers potentially is a powerful way to improve science instruction, since each teacher directly affects up to six or seven classes of students during each semester or quarter, considerably more than elementary school teachers.

The landmark report *Taking Science to School: Learning and Teaching Science in Grades K–8*, produced by the National Research Council in 2007, concludes that "well-designed opportunities for teacher learning can produce desired changes in their classroom practices, can enhance their capacity for continued learning and professional growth, and can, in turn contribute to improvements in student learning" (Duschl, Schweingruber, and Shouse 2007, pp. 306–07). The most successful features of professional development described in the literature include a focus on content; teacher curricula grounded in classroom experiences and linked to standards-based, high-quality student curricula; and a process that offers teachers opportunities for professional dialogue and critical reflection (Cohen and Hill 2000, 2001; Desimone et al. 2002; Garet et al. 2001; Kennedy 1998; Knapp, McCaffrey, and Swanson 2003; Little 2006; National Staff Development Council 2001; Weiss et al. 1999; Wilson and Berne 1999).

Embodying these characteristics, the Making Sense of SCIENCE™ approach focuses on developing teachers' pedagogical and content knowledge. The model is based on the premise that, to develop this specialized knowledge, teachers must have opportunities to learn science content knowledge in combination with analysis of student thinking about that content and they need instructional strategies for helping students learn that content (Duschl, Schweingruber, and Shouse 2007; Shinohara, Daehler, and Heller 2004; Shymansky and Matthews 1993; Van Driel, Verloop, and De Vos 1998). Previous empirical studies provide consistent evidence that the Making Sense of SCIENCE™ model is effective for improving student science achievement in elementary school (Heller, Daehler, and Shinohara 2003; Heller and Kaskowitz 2004). To date, however, the effectiveness of the program for middle school science achievement has not been examined.

Some have argued that most school districts in the United States lack coherent, effective professional development programs, site-based expertise, and science-savvy staff developers to provide such programs (Little 2006; Duschl, Schweingruber, and Shouse 2007). Given the strong need for effective professional development programs that address teachers' content knowledge of science, the 2007 National Research Council report called for comprehensive professional development programs that are "conceived of, designed, and implemented as a coordinated system" to support students' attainment of high standards (Duschl, Schweingruber, and Shouse 2007, p. 347).

Training teachers in Making Sense of SCIENCE™

A course from the WestEd Making Sense of SCIENCE™ series was chosen for this study because it had a history of promising empirical evidence of effectiveness and an unusual combination of features, including opportunities for teachers to learn science content knowledge along with analysis of student thinking about that content and analysis of instructional strategies for helping students learn the content. Most other professional development programs deal with just one or two of these areas (for example, science content or teaching), leaving teachers the task of knitting together the information they most need to do their jobs well. Making Sense of SCIENCE™ courses also focus on science literacy by helping teachers and their students build important skills for reading and making sense of science texts.

The course includes numerous key features of professional development that have been associated with increasing student achievement (Birman, Desimone, Porter, and Garet 2000; Desimone 2009): (a) in-depth focus on science content; (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, 24 hours over five days; and (e) a process of collective participation during which teachers engage in professional discourse and critical reflection. Although sustained involvement in professional development activities has been found to be associated with better outcomes, the evidence regarding the necessity of extended school-year activities is not conclusive (Wayne, Yoon, Zhu, Cronen, and Garet 2008), and previous research on five-day Making Sense of SCIENCE™ intensive workshops has found strong effects for teachers and students (e.g., Heller, Daehler, and Shinohara 2003, 2011). Similarly, Desimone (2009) states, "Research has not indicated an

exact 'tipping point' for duration but shows support for activities that are spread over a semester (or intense summer institutes with follow-up during the semester) and include 20 hours or more of contact time" (p. 184).

The WestEd courses are designed around two main components—hands-on science investigations and discussions of narrative teaching cases (Daehler and Shinohara 2001). They were written by classroom teachers and field tested with ethnically, culturally, socioeconomically, and linguistically varied groups of students and teachers from across the U.S. The case materials are drawn from actual classroom episodes and contain descriptions of instructional activities, student work including examples of common but incorrect ways students think about concepts, student-teacher dialogue, and teacher thinking and behaviors. The hands-on science investigations conducted by students, as described in the narrative cases, parallel the science investigations done by teachers in each session, thus building on research findings that teachers' knowledge grows when teachers encounter subject content through school curricula (Cohen and Hill 2001; Saxe, Gearhart and Nasir 2001). In addition to these two components, language and literacy activities support students' science reading and discussion skills; help students make sense of the science; and help students, particularly English language learners, develop their academic language proficiency.

Making Sense of SCIENCE™ courses provide firsthand experiences for teachers in ways of learning science that research suggests are effective for all students and especially for English language learners. English language learners can benefit greatly from inquiry-based science instruction (Hewson, Kahle, Scantlebury, and 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 and Fradd 2001).

The professional development intervention was implemented regionally, with local facilitators leading the course for local teachers at each of six research sites. The five course sessions were sequenced so that the science topics (for example, speed, velocity, acceleration, and balanced and unbalanced forces) built on one another. The corresponding science language issues and strategies for supporting student learning and language development were unveiled incrementally over the sessions.

Measuring the impact of Making Sense of SCIENCE™ on students and teachers

This study was an experimental trial designed to test the effects of a Making Sense of SCIENCE™ course on force and motion on grade 8 students' knowledge of course content, as measured by the Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion (http://www.horizon-research.com/atlast/; Smith and Banilower 2006a) Impacts on these outcomes were estimated for all grade 8 students in the study sample and for the subsample of English language learners. The study also estimated program effects on teachers' content knowledge of force and motion, as measured by the ATLAST Test of Force and Motion for Teachers (http://www.horizon-research.com/atlast/; Smith and Banilower 2006b) and by their self-reported confidence in teaching force and motion.

The study sample included 181 teachers from 137 schools in 55 districts who were randomly assigned to an intervention or control group (90 to intervention and 91 to control). The trial was conducted at six regional sites, five in California and one in Arizona. Each site was comprised of multiple school districts in the region from which teachers were drawn, and the intervention was implemented once at each of these six sites.

The study was conducted from spring 2009 through spring 2010. Outcomes were measured for teachers during both the 2008/09 and 2009/10 school years and for students during the 2009/10 school year. Teachers in the intervention group received a 24-hour Making Sense of SCIENCE^{TM} professional development course on force and motion in summer 2009. Intervention group teachers did not receive additional Making Sense of SCIENCE^{TM} professional development or support during the school year.

About 72 percent of the original 181 teachers completed the study and provided survey and test data (77 percent of the intervention group teachers and 70 percent of the control group teachers). Nine intervention group teachers (10 percent) and 10 control group teachers (11 percent) dropped out; 29 teachers were not retained for reasons outside of their control. The 133 teachers who were retained in the analytic sample after attrition came from 102 schools in more than 40 districts. Research sites after attrition included 2–10 districts and 13–21 schools.

At each school, proctors administered student science tests, following a detailed testing protocol provided by the research team. Consistent with common practice for the administration of standardized tests in schools, test proctors were professional staff members who were not directly involved in the classroom being studied (counselors, aides, administrators, other teachers).

Regional site coordinators administered teacher science tests and surveys to both intervention and control group teachers in regional project meetings in winter/spring 2009, before random assignment to condition, and in fall/winter 2010, after teachers completed teaching the force and motion unit in their classes and students had taken their posttests. Site coordinators were provided with detailed test administration instructions.

Multilevel regression models that accounted for the nesting of students within teachers and teachers within sampling blocks were used to estimate the impact of the professional development. When warranted, statistical significance levels of the impact estimates were adjusted to account for multiple comparisons within domains. To deal with item-level missing values in constructed measures, the research team created total scale scores by averaging items with non-missing values. It used the missing indicator method to account for missing values in the impact analysis models (White and Thompson 2005). Then, the analytic models included categorical variables to denote whether or not the value of a particular variable was missing.

Study findings

Results for the primary confirmatory analyses indicate that after adjusting for multiple comparisons, there were no statistically significant differences between the test results on science content of students in intervention group classrooms and students in control group classrooms. Intervention group students in neither the full sample (effect size = 0.11) nor the English language learner subsample (effect size = 0.31) scored significantly higher on the ATLAST Test of Force and Motion than did their control group counterparts. Similarly, intervention group students in neither the full sample (effect size = 0.03) nor the English language learner subsample (effect size =0 .09) scored higher on the physical science reporting clusters of the California Standards Test than did their control group counterparts.

Results for the intermediate confirmatory analyses indicate that after adjusting for multiple comparisons, teachers who received the professional development course outscored their control group counterparts on the ATLAST Test of Force and Motion for Teachers (effect size = 0.38), as well as on their ratings of confidence in their ability to teach force and motion (effect size = 0.49).

With one exception, the study findings were not sensitive to variations in specification of the estimation models. The exception is that, for teacher content knowledge, inclusion of the pretest in the impact analysis model (basic model plus pretest) decreased the point estimate from 9.8 to 6.1 and the effect size from 0.61 to 0.38.

In exploratory analyses, the study investigated whether there were differential impacts on student and teacher content knowledge outcomes across the six research sites. The estimated impacts were most pronounced at two of the six sites. For the full sample of students, point estimates for student and teacher content knowledge of force and motion followed exactly the same rank order at all sites.

Limitations

There are three main limitations of this study. First, there was high sample attrition: 48 of the 181 teachers who were randomly assigned to intervention and control groups left the study before data collection was completed. However, there is no evidence that attrition resulted in significant differences at the baseline between the intervention and control samples used in the analysis.

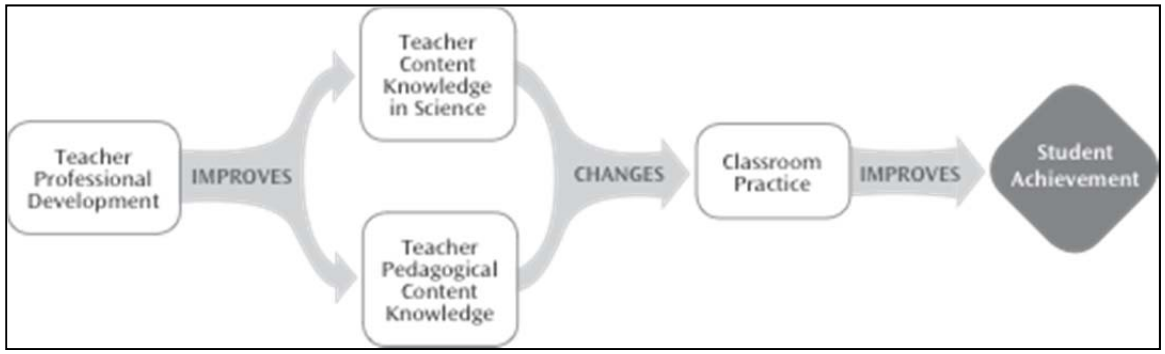
Second, the study did not include analyses of classroom implementation of course-related practices. As a result it is not possible to infer whether the lack of student effects is due to a failure of treatment group teachers to modify classroom practices or a failure of modified practices to affect student outcomes. Third, the findings are based on volunteer teachers and students whose parents provided consent. It is possible that the findings would have been different had teachers been required to participate in the intervention, and all students been tested.

Chapter 1. Overview of study

This study evaluated an approach to professional development for middle school science teachers. The study is a cluster-randomized controlled trial designed to test the effectiveness of a Making Sense of SCIENCE™ course on force and motion (Daehler, Shinohara, and Folsom 2011). The study compares outcomes for students of teachers who took the course with outcomes for students of teachers in a control group that included no science professional development beyond that ordinarily received. Outcomes for teachers were also evaluated. The research was conducted at six regional sites, five in California and one in Arizona, by Heller Research Associates (HRA), an evaluation firm external to WestEd, the REL West contractor and developer of the intervention.

Theoretical models of effective teacher professional development share a fundamental assumption that there is a cascade of influences from features of the professional development to immediate impact on teacher knowledge, intermediate impacts on classroom instruction, and more distal effects on student achievement (see Figure 1.1) (Cohen and Hill, 2000; Desimone, 2009; Heller, Daehler, and Shinohara, 2003; Scher and Reilly, 2009; Weiss and Miller, 2006). As summarized in the conclusion of the landmark National Research Council report, *Taking Science to School: Learning and Teaching Science in Grades K–8*, "…well-designed opportunities for teacher learning can produce desired changes in their classroom practices … and can in turn contribute to improvements in student learning" (Duschl, Schweingruber, and Shouse, 2007, pp. 306–07). A growing body of empirical evidence supports this claim that teacher professional development can strengthen student achievement (e.g., Blank, de las Alas, and Smith, 2007; Fennema, Carpenter, Franke, Levi, Jacobs, and Empson, 1996; Franke, Carpenter, Levi, and Fennema, 2001; Saxe, Gearhart and Nasir, 2001), and there is increasing consensus on key characteristics of effective professional development (Desimone, 2009; Yoon, Duncan, Lee, Scarloss, and Shapley, 2007).

Figure 1.1 Making Sense of SCIENCE™ theory of action



Source: Adapted from Horizon Research's ATLAST Theory of Action model http://www.horizon-research.com/atlast

There is, however, little evidence about the impact of specific professional development features on teacher knowledge or student achievement (Wayne, Yoon, Zhu, Cronen, and Garet 2008), or about relationships between particular aspects of teacher change and student outcomes (Borko 2004; Desimone, Porter, Garet, Yoon, and Birman 2002; Fishman, Marx, Best, and Tal 2003; Garet, Porter, Desimone, Birman, and Yoon 2001; Scher and Reilly

2009). Furthermore, the literature to date largely demonstrates the efficacy of professional development interventions that are delivered by the developers of the inservice courses to relatively small numbers of teachers and schools. Effectiveness trials have been called for to test delivery of interventions by multiple trainers in a range of typical settings for which the interventions are designed, as a critical step toward scaling up effective practices (Borko, 2004; Wayne, et al., 2008). This study addresses some of these knowledge gaps by estimating the effects of a particular professional development program on outcomes for students and teachers using a large-scale experimental design study. The study used a randomized experimental design, as has been encouraged in educational research (Boruch, DeMoya, and Snyder, 2002; Jacob, Zhu, and Bloom, 2010; Slavin, 2002).

Teacher courses developed by the Understanding Science for Teaching program, including Making Sense of SCIENCE™, are intended to improve students' science achievement, including that of low-performing students and English language learners, by strengthening their teachers' science content knowledge and knowledge for teaching that science. In a national randomized experiment and numerous field tests (Heller, Daehler, and Shinohara 2003; Heller et al. 2010), Making Sense of SCIENCE™ courses have been shown to increase elementary school teachers' content knowledge and student achievement. The effectiveness of the courses has not been examined for middle school teachers and students.

The professional development tested in this study has the potential to significantly improve methods for preparing novice and experienced teachers alike. The ultimate potential long-term contribution of this work is greater nationwide gains in middle school students' science achievement, resulting from widely available, low-cost staff development courses that enhance teachers' science content knowledge and improve their teaching practices.

Need for better preparation of science teachers

The world of work requires skills learned in science, such as deep critical thinking, inquiry, problem solving, and teamwork. Science education is important for closing the skills gaps and responding to the labor needs and shortages in the workforce (Partnership for 21st Century Skills 2008) , particularly in light of the job growth in professional occupations, such as health care and education, and in technical fields, such as computing (Terrell 2007).

Many states have responded by setting high standards for students' science learning. For students to attain these standards, their teachers not only need a strong grasp of the subject matter, they must also know "how to organize, sequence, and present the content to cater to the diverse interests and abilities of the students" (Barnett and Hodson 2001, p. 432). Teachers are a dominant factor affecting student academic achievement (Duschl, Schweingruber, and Shouse 2007; Hill, Rowan, and Ball 2005). Teachers with strong content knowledge and science-specific pedagogical knowledge are more likely to anticipate difficulties students may encounter, encourage students to discuss the content and think about applications, and use accurate representations (Carlsen 1991, 1993; Hashweh 1987).

One study reported that more than two-thirds of middle school science teachers in the United States have inadequate science preparation (Fulp 2002). "Out-of-field" teaching is widespread and stands to increase as many veteran science teachers retire. For example, Fulp (2002) reported that only 28 percent of science teachers in grades 6–8 have an undergraduate degree

in science (Fulp 2002). Quality professional development for middle school teachers may be especially important, because each teacher directly affects up to six or seven classes of students each term, considerably more than elementary school teachers.

There is a significant disjuncture between what is known about quality professional development and what is available to districts, especially those with poor student achievement and inadequate teacher preparation. Many districts in the United States apparently lack coherent, effective professional development programs, site-based expertise, and science-savvy staff developers to provide effective programs (Little 2006; Duschl, Schweingruber, and Shouse 2007).

Rationale for choice of course on force and motion

We chose to study the Making Sense of SCIENCE^{TM} course on force and motion for three reasons. First, it is well documented that physical science is an especially problematic content area for middle school science teachers (Fulp 2002). Nearly half of all middle school physical science classes are taught by teachers who lack in-depth preparation in any science (Fulp 2002), and 74 percent of more than 5,700 middle school science teachers surveyed in the 2000 National Survey of Science and Mathematics Education had two semesters or less of coursework in physical science (Weiss et al. 2001).

Second, the topic of force and motion is a prominent topic in kit-based science curricula in grades 6–8: it is one of nine Full Option Science System (FOSS) middle school science modules (Delta Education, 2010), one of five Science/Technology/Engineering/Mathematics Curriculum Integration Program (STEM-CIP) middle school science modules (Hawker Brownlow, 2010), and one of eight Science and Technology Concepts (STC) middle school science modules (Carolina Curriculum for Science and Math, 2010). In California the topic constitutes one-third of the science curriculum for grade 8 students.

Third, the topic is covered in 35–50 percent of the chapters in the most frequently used physical science textbooks (Fulp 2002), but misconceptions about it on the part of students and teachers are well documented (American Association for the Advancement of Science 1993; Driver, Guesne, and Tiberghien 1985; Hapkiewicz 1999). Given the centrality of this topic in the middle grades, students have the potential to make sizable gains in their overall science achievement scores if they are taught by teachers who are better prepared to teach this topic.

Addressing the needs of students with limited English proficiency

Science achievement for English language learners lags well behind that for native English speakers in the United States (Torres and Zeidler 2002). Both states in this study—California and Arizona—have high percentages of English language learners. During the 2008/09 school year, more than 1.5 million students enrolled in California public schools (25 percent of all public school students) and close to 59,000 in Arizona (10 percent of all public school students) were designated English language learners. Among grade 8 students who took the 2009 California Standards Test, only 18 percent of English language learners scored "Proficient" or higher on the science portion of the test, compared with 56 percent of all grade 8 students (California Department of Education 2011a). Among grade 8 students in Arizona

who took the 2009 Arizona Instrument to Measure Standards, only 6 percent of English language learners scored "Meets" or higher on the science portion of the test, compared with 56 percent of all grade 8 students (Arizona Department of Education 2010).

Nearly all middle school students are challenged by the density of science textbooks; the challenge is particularly great for English language learners (ELLs). "To keep from falling behind their English-speaking peers in academic content areas, such as science, ELLs need to develop English language and literacy skills in the context of subject area instruction" (Lee 2005, p. 492). To support the science achievement of English language learners, teachers need strong and integrated knowledge of the science and knowledge of English language and literacy development.

The Making Sense of SCIENCE™ professional development is designed to build this particular combination of teacher knowledge. It includes an intensive science content component along with activities to help teachers support students' reading, writing, and speaking in the languages and culture of science as a means to help students make sense of the material and develop academic language proficiency. A full quarter of the program focuses teachers' attention on identifying and evaluating literacy supports that guide learning. For example, the course is intended to help teachers understand that, in order to lead successful discussions about science ideas, they need to make data public, visual, and manipulable, so that students can discuss data sets, make comparisons, and draw conclusions. Teachers also practice and are expected to become fluent in using the representations most commonly used to organize and display data in different science disciplines, including number lines, graphs, tables, and equations. Making Sense of SCIENCE™ professional development is intended to help teachers gain a clear understanding of the purpose and utility of different representations, so that they can use them more purposefully.

Making Sense of SCIENCE™ courses are intended to prepare teachers to improve all students' science achievement and academic literacy skills. To accomplish this, they model and provide firsthand experiences for teachers in ways of learning science that research suggests are effective for all students and especially for English language learners. The courses include features that implement the Five Standards for Effective Pedagogy for students whose ability to reach their potential is challenged by language or cultural barriers (Tharp, Estrada, Dalton, and Yamauchi 2000) developed by the Center for Research on Education, Diversity & Excellence of the Graduate School of Education (CREDE) at the University of California, Berkeley (http://gse.berkeley.edu/research/credearchive/standards/standards.html) (see table 1.1).

Table 1.1. Course features corresponding to CREDE standards for effective pedagogy for students whose ability to reach their potential is challenged by language or cultural barriers

Strategy	Making Sense of SCIENCE™ features
• Teachers and Students Working Together: Use instructional group activities in which students and teacher work together to create a product or idea.	Collaborative group science investigations and sense-making discussions
• Developing Language and Literacy Skills across All Curricula: Apply literacy strategies and develop language competence in all subject areas.	Reading, writing, and speaking activities in science along with interpreting diagrams, graphs, and tables to develop academic language proficiency
• Connecting Lessons to Students' Lives: Contextualize teaching and curriculum in students' existing experiences in home, community, and school.	Hands-on activities based on natural phenomena that students experience in class
• Engaging Students with Challenging Lessons: Maintain challenging standards for student performance; design activities to advance understanding to more complex levels.	Instructional tasks focused on making meaning of complex science ideas
• Emphasizing Dialogue over Lectures: Instruct through teacher-student dialogue, especially academic, goal-directed, small-group conversations (known as instructional conversations), rather than lecture.	Small-group opportunities for developing English proficiency through authentic communication about science ideas and observations

Source: Strategy column is drawn from the Five Standards for Effective Pedagogy for students whose ability to reach their potential is challenged by language or cultural barriers (Tharp, Estrada, Dalton, and Yamauchi 2000) developed by the Center for Research on Education, Diversity & Excellence of the Graduate School of Education (CREDE) at the University of California, Berkeley (http://gse.berkeley.edu/research/credearchive/standards/standards.html).

Making Sense of SCIENCE™ courses provide firsthand experiences for teachers in ways of learning science that research suggests are effective for all students and especially for English language learners. English language learners can benefit greatly from inquiry-based science instruction (Hewson, Kahle, Scantlebury, and 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 and Fradd 2001).

To support teachers in capitalizing on what they learn, Making Sense of SCIENCE™ provides them with opportunities to plan how they might modify their instruction by incorporating literacy supports and attending to English language learners' needs in their classrooms. For example, teachers plan discussion sequences with clear participation structures, with the intention of helping their English language learners learn "the rules of the game" so that they can more actively and successfully participate in scientific discourse. Teachers plan hands-on learning in small groups to allow students to rehearse science language and ideas before presenting them in a higher-risk setting. Teachers plan ways of

making data from investigations accessible by incorporating objects from life outside of school into their classroom discussions and writing assignments for students.

Overview of the intervention

The intervention implemented in this study—a Making Sense of SCIENCE™ professional development course for grade 8 science teachers—embodies characteristics described in the research literature on effective programs. The landmark report *Taking Science to School: Learning and Teaching Science in Grades K–8*, produced by the National Research Council in 2007, concludes that "well-designed opportunities for teacher learning can produce desired changes in their classroom practices, can enhance their capacity for continued learning and professional growth, and can in turn contribute to improvements in student learning" (Duschl, Schweingruber, and Shouse 2007, pp. 306–07). The most successful features of professional development described in the literature include a focus on content; teacher curricula grounded in classroom experiences and linked to standards-based, high-quality student curricula; and a process that offers teachers opportunities for professional dialogue and critical reflection (Cohen and Hill 2000, 2001; Desimone et al. 2002; Garet et al. 2001; Kennedy 1998; Knapp, McCaffrey, and Swanson 2003; Little 2006; National Staff Development Council 2001; Weiss et al. 1999; Wilson and Berne 1999).

In the context of the strong need for effective professional development programs that address teachers' content knowledge of science, the 2007 National Research Council report called for comprehensive professional development programs that are "conceived of, designed, and implemented as a coordinated system" to support students' attainment of high standards (Duschl, Schweingruber, and Shouse 2007, p. 347). The Making Sense of SCIENCE™ professional development courses offer just this kind of program. A course from the WestEd Making Sense of SCIENCE™ series was chosen for this study because it had a history of promising empirical evidence of effectiveness and an unusual combination of features, including opportunities for teachers to learn science content knowledge along with analysis of student thinking about that content and analysis of instructional strategies for helping students learn the content. Most other professional development programs deal with just one or two of these areas (for example, science content or teaching), leaving teachers with the task of knitting together the information they most need to do their jobs well. Making Sense of SCIENCE™ courses focus on literacy by helping teachers and their students build important skills for reading and make sense of science texts. This unique component is one reason why the Making Sense of SCIENCE™ courses have the potential to be particularly effective with English language learners.

The Making Sense of SCIENCE™ approach focuses on developing teachers' pedagogical content knowledge. The model is based on the premise that, to develop this specialized knowledge, teachers must have opportunities to learn science content knowledge in combination with analysis of student thinking about that content and analysis of instructional strategies for helping students learn that content (Duschl, Schweingruber, and Shouse 2007; Shinohara, Daehler, and Heller 2004; Shymansky and Matthews 1993; Van Driel, Verloop, and De Vos 1998). Previous empirical studies provide evidence that this model is effective for improving student science achievement (Heller, Daehler, and Shinohara 2003; Heller and Kaskowitz 2004).

The course includes numerous key features of professional development that have been associated with increasing student achievement (Birman, Desimone, Porter, and Garet 2000; Desimone 2009): (a) in-depth focus on science content; (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, 24 hours over five days; and (e) a process of collective participation during which teachers engage in professional discourse and critical reflection. Although sustained involvement in professional development activities has been found to be associated with better outcomes, the evidence regarding the necessity of extended school-year activities is not conclusive (Wayne, Yoon, Zhu, Cronen, and Garet 2008), and previous research on five-day Making Sense of SCIENCE™ intensive workshops has found strong effects for teachers and students (e.g., Heller, Daehler, and Shinohara 2003, 2011). Similarly, Desimone (2009) states, "research has not indicated an exact 'tipping point' for duration but shows support for activities that are spread over a semester (or intense summer institutes with follow-up during the semester) and include 20 hours or more of contact time" (p. 184).

Structure of the intervention

Making Sense of SCIENCE™ draws on research on adult learning and cognitive psychology. Its course structure is designed to move teachers through learning about key science concepts, literacy supports, classroom practices, and students' science ideas. Courses have four main components:

- Hands-on science investigations engage teachers in core content dilemmas described in accompanying written teaching cases. The investigations parallel those of students in the teaching cases, in the context of commonly used, standards-based curricula.
- Language and literacy activities are intended to teach teachers how to more effectively support students' science reading and discussion skills; help students make sense of the science; and help students, particularly English language learners, develop their academic language proficiency.
- Case discussions engage teachers in examining detailed instructional scenarios. The materials, written by classroom teachers, contain student work, student/teacher dialogue, context information, and discussions of teacher thinking and behavior. Teachers examine student thinking and critically analyze instruction presented in the cases.
- Classroom connections provide opportunities for teachers to read about, reflect on, and discuss key science and literacy concepts and consider how these concepts pertain to their own work with students.

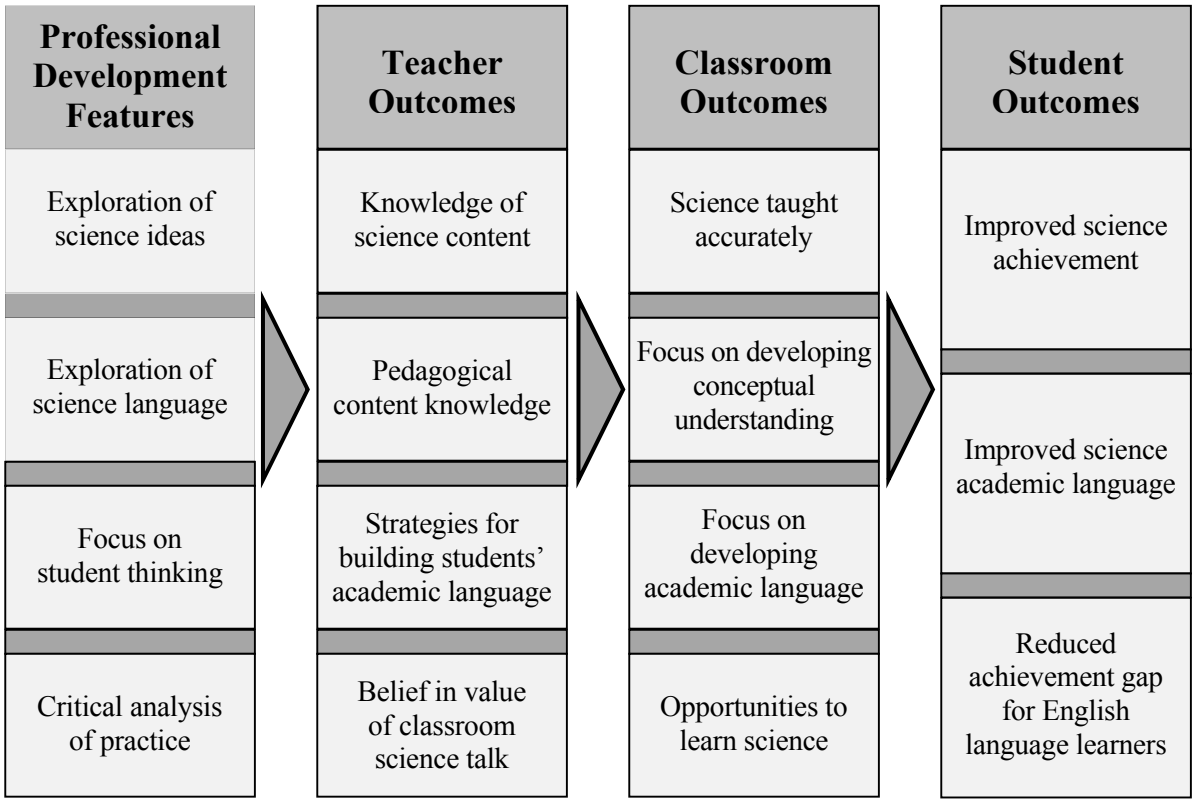
The materials for each course include a facilitator guide that provides detailed yet flexible procedures; in-depth background information (for example, descriptions of the underlying science and common but incorrect ideas teachers have); guiding questions and charts for each whole-group discussion; and other tips for leading a successful course. An accompanying

teacher book presents all the materials teachers need to teach a course, including teaching cases, handouts, and session reviews that summarize the key concepts and outcomes and feature illustrations of common but incorrect ways students think about related concepts.

Professional development logic model

The logic model motivating this approach describes the cascade of influences connecting teachers' experiences in Making Sense of SCIENCE™ courses to student outcomes (figure 1.2). The theory of action posits that professional development that is situated in an environment of collaborative inquiry—one that is rich in talk about scientific meanings, in conjunction with a focus on student thinking and critical analysis of practice—leads to increases in teachers' science content and pedagogical content knowledge, along with important shifts in teachers' strategies for supporting students' literacy needs and in teachers' beliefs about the role of literacy in science classrooms. These outcomes for teachers result in changes in classroom practices, such as increased accuracy of science representations and explanations, a focus on conceptual understanding, greater opportunity for students to read and write to learn, and explicit development of academic language. Classroom changes ultimately produce improvements in student achievement, along with increased development of all students' literacy abilities and reduced achievement gaps for low-performing students and English language learners.

Figure 1.2 Making Sense of SCIENCE™ logic model



Source: Author.

Previous evidence on the effects of Making Sense of SCIENCE™

Over the past decade, a series of increasingly rigorous quasi-experimental and experimental studies of the Making Sense of SCIENCE™ professional development model have documented its effects on the science achievement of high-needs K–8 students, including English language learners. Statistically significant differences were found favoring intervention teachers and students on measures of science content knowledge in pilot tests and national field tests (Heller, Daehler, and Shinohara 2003, 2011; Heller et al. 2010). Project teachers showed significant gains of more than one standard deviation on tests of content knowledge about electricity and magnetism (Heller and Kaskowitz, 2004), and important changes in pedagogical content knowledge as demonstrated through in-depth assessment interviews requiring reasoning about student work and instruction (Heller, Daehler, and Shinohara 2003, 2011; Heller et al. 2010). In every field test, statistically significant differences in measures of science content knowledge were found favoring intervention group teachers and students with effect size statistics for teachers ranging from just under one standard deviation unit difference between posttest and pretest means ($ES = 0.7$) to more than one standard deviation difference ($ES = 1.3$) and effect sizes from 0.4 and 0.8 for students (Heller, Daehler, and Shinohara 2003, 2011; Heller et al. 2010). The data from a large randomized experiment in six states offered strong evidence of the model's impact on elementary school students' achievement across states; districts of varying sizes; and diverse urban student populations, with both native English speakers and English language learners and a range of socioeconomic backgrounds, with effect sizes 0.5-0.8 for students (Heller, Daehler, and Shinohara 2011; Heller, Daehler, Shinohara, and Kaskowitz, 2004). Collectively, these data provide strong evidence of the internal validity of the professional development model.

One of the most rigorous tests of the Making Sense of SCIENCE™ model was conducted by researchers from the University of California, Berkeley, and Heller Research Associates, with support from the National Science Foundation (Heller et al. 2010). They conducted a cluster-randomized experiment over a two-year period (2007–09) to test the Making Sense of SCIENCE™ model in eight sites across the United States that included 49 districts and more than 260 elementary school teachers. The nearly 7,000 students in the study came largely from underserved populations, including some classrooms in which 100 percent of students were eligible for free or reduced-price meals, and up to 65 percent were English language learners.

The intervention was a Making Sense of SCIENCE™ professional development course on electric circuits. Tests of content knowledge of electric circuits were administered to all teachers at the beginning and end of the 2007/08 school year and a year later; students were tested before and after the classroom units on electric circuits during the 2007/08 and 2008/09 school years. Because no off-the-shelf tests were available, the teacher and student tests were developed by the research staff and validated for use in previous evaluations of the Making Sense of SCIENCE™ course on electric circuits (Heller et al. 2010). These tests were aligned with the Understanding Science for Teaching project content framework, which specified the targets of instruction based on National Science Education Standards (National Research Council 1996); Benchmarks for Science Literacy (American Association for the Advancement of Science 1993); a host of state content standards; and frequently used kit-

based student science curricula, such as Full Option Science System (FOSS) (Delta Education, 2010), Science and Technology Concepts (STC) (Carolina Curriculum for Science and Math, 2010), and Curriculum Integration Program (STEM-CIP) (Hawker Brownlow, 2010). The tests included questions reflecting the format and content of questions in the Trends in International Mathematics and Science Study (U.S. Department of Education 2004) and the National Assessment of Educational Progress. Cronbach's alpha coefficients were determined to be 0.87 for the student tests and 0.90 for the teacher tests.

A teaching background survey provided data on all teachers' professional experience and perspectives on science teaching. A randomly selected subsample of teachers participated in pre- and post-professional development interviews designed to elicit their pedagogical content knowledge. Teachers were also observed and videotaped twice while teaching lessons on electric circuits. Data were collected in two rounds of professional development course implementation.

Results showed that a single Making Sense of SCIENCE^{TM} course produced exceptional gains in elementary school teachers' content knowledge about electric circuits. Teachers who took the course increased the percentage of items they answered correctly on a knowledge test by 21.0 percentage points, on average, compared to an increase of 1.4 percentage points for control group teachers ($p < 0.001$, effect size = 1.8).[2] Significant treatment effects were also found at the student level for content knowledge. The percentage of items answered correctly by students in treatment teachers' classrooms increased by 18.4 percentage points, compared with 13.3 percentage points for students in control group teachers' classrooms ($p < 0.001$, effect size = 0.36) (Heller et al. 2010). Unadjusted mean gains for student subgroups classified at different levels of English language proficiency show that the greater score increases for students of intervention teachers also occurred for all subgroups of English proficiency, with mean gains of 15.5 percentage points for intervention students with little or no English, compared with control student mean gains of 6.0 percentage points ($p < 0.001$, effect size = 0.7), and 17.0 percentage points for intermediate English proficient students, compared with control student means of 9.2 percentage points ($p < 0.001$, effect size = 1.3). Furthermore, treatment effects for both teachers and students were maintained a full year later, with students of intervention group teachers showing gain scores that were significantly greater than those of students of control group teachers (Heller, Daehler, and Shinohara 2011; Heller et al. 2010). Qualitative research also documents differences in the teaching practices, pedagogical reasoning, and pedagogical content knowledge of intervention and control group teachers

The findings from the randomized controlled trial were preceded by five years of quasi-experimental evaluation studies beginning in 2000 that identified positive teacher and student outcomes of various Making Sense of SCIENCE^{TM} courses for elementary and middle school teachers. Although the non-experimental evidence did not allow definitive conclusions to be drawn, the pattern of quantitative and qualitative findings suggests that gains were the

[2] These numbers represent the most conservative measures of effect size with 95 percent confidence intervals based on the standard error of the difference in mean change in scores between the intervention and control groups. Effect size was computed as the hierarchical linear model coefficient divided by the pooled standard deviations of the teacher gains.

result of teachers' participation in the Making Sense of SCIENCE™ courses. Findings of the previous study (Heller et al. 2010) include the following:

- For teachers at both the elementary and middle school levels, differences between teachers' mean pre- and post-course scores on science tests were statistically significant in every study of Making Sense of SCIENCE™ courses, with effect sizes of 0.44–1.09.
- At the elementary school level, statistically significant differences favoring students in the intervention group were found between the adjusted posttest mean for students of teachers who participated in Making Sense of SCIENCE™ courses (n = 123) and the adjusted posttest mean for the comparison groups (n = 84) after controlling for pretest differences (effect size = 0.84).
- English language learners in the intervention group (n = 97) made gains that were statistically significant, raising their scores by 0.95 standard deviation more than English language learners in the control group (n = 57).
- Students of all ability levels showed significant gains, with the greatest increase among low-performing students of intervention group teachers (effect size = 1.02).

Overall, these studies provide strong experimental evidence of the effectiveness of Making Sense of SCIENCE™ at the elementary school level and moderate quasi-experimental evidence of its effectiveness at the middle school level. Although the same professional development model is incorporated in Making Sense of SCIENCE™ courses at the two levels, it would be premature to conclude based on previous studies that the program's middle school courses are effective. The many contextual and curricular differences between elementary and middle school science warrant more rigorous investigation of the program for higher-grade teachers and students.

Research questions

This study was designed to test the effects of the Making Sense of SCIENCE™ model of professional development by closely examining one grade 8 course (on force and motion) that embodies that approach. The study estimated the effects of the program on both students, including English language learners and teachers.

Confirmatory research questions

Primary confirmatory questions: student outcomes. The study examined two primary confirmatory questions:

1. What is the impact of the Making Sense of SCIENCE™ professional development course on students' content knowledge of force and motion and of physical science more generally?

   Hypothesis 1a: Making Sense of SCIENCE™ professional development increases students' content knowledge of force and motion.

Hypothesis 1b: Making Sense of SCIENCE™ professional development increases students' content knowledge of physical science more generally.

2. What is the impact of the Making Sense of SCIENCE™ professional development course on English language learners' content knowledge of force and motion and of physical science more generally?

   Hypothesis 2a: Making Sense of SCIENCE™ professional development increases English language learners' content knowledge of force and motion.

   Hypothesis 2b: Making Sense of SCIENCE™ professional development increases English language learners' content knowledge of physical science more generally.

Intermediate confirmatory questions: teacher outcomes. The theory of action that links the Making Sense of SCIENCE™ professional development course to students' academic skills and knowledge holds that the intervention increases teachers' knowledge of science content and instruction while helping teachers develop targeted strategies for eliciting student ideas and strengthening their science language abilities. The study posits that these outcomes will lead to changes in classroom practices that ultimately improve student achievement. To examine part of this logic model, the study examined the impact of the professional development course on teachers' content knowledge and self-reports of confidence in their ability to teach force and motion. Specifically, it examined the following questions:

3. What is the impact of the Making Sense of SCIENCE™ professional development course on teachers' content knowledge of force and motion?

   Hypothesis 3: Making Sense of SCIENCE™ professional development increases teachers' content knowledge of force and motion.

4. What is the impact of the Making Sense of SCIENCE™ professional development course on teachers' confidence in their ability to teach force and motion?

   Hypothesis 4: Making Sense of SCIENCE™ professional development increases teachers' confidence in their ability to teach force and motion.

Exploratory research questions

Exploratory analyses investigated whether the impacts of the intervention on teacher and student outcomes differed across the six regional sites, whether the pattern of differences in impact across sites varied for teacher and student outcomes, and the extent to which program impacts on student outcomes were mediated by teacher content knowledge.

The study addressed the following exploratory questions for each hypothesis in the research plan, for both the full sample and the subsample of English language learners.

Exploratory research question: student outcomes. The study examined two exploratory research questions concerning student outcomes:

1. Do the impacts of the Making Sense of SCIENCE™ professional development course on students' content knowledge of force and motion vary by site?
2. Do the impacts of the Making Sense of SCIENCE™ professional development course on English language learners' content knowledge of force and motion vary by site?

Exploratory research question: teacher outcomes. The study examined one exploratory research question concerning teacher outcomes:

3. Do the impacts of the Making Sense of SCIENCE^{TM} professional development course on teachers' content knowledge of force and motion vary by site?

Measures of key outcomes

Primary student outcomes were measured with instruments that capture student content knowledge of force and motion and of physical science more generally (table 1.2). Intermediate teacher outcomes included content knowledge of force and motion and confidence in teaching abilities.

Table 1.2. Key outcome variables and data collection measures, by outcome domain

Outcome variable	*Measure*
Student content knowledge of force and motion	Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion for Students (Smith and Banilower 2006a, 2006b)
Student content knowledge of physical science more generally	California Standards Test reporting clusters on motion (8 items) and forces, density, and buoyancy (13 items) (California Department of Education, 2011b)
Teacher content knowledge of force and motion	ATLAST Test of Force and Motion for Teachers (Smith and Banilower 2006a, 2006b)
Confidence in ability to teach	Teacher survey administered as part of this study

Structure of report

Chapter 2 describes the study design, including recruitment of teachers and students, random assignment to intervention and control groups, collection of data, selection of analytic study samples, and methods of data analysis. It also examines sample attrition and baseline equivalence at both the teacher and student levels. Chapter 3 describes the intervention. Chapter 4 reports the results of the impact analyses for the experimental findings. Chapter 5 reports the results of the exploratory analyses examining differential site-level impact. Chapter 6 summarizes the findings and explores what the results may mean to educators, policymakers, and researchers.

Chapter 2. Research design and methods

The goal of this study was to evaluate the efficacy of a Making Sense of SCIENCE™ professional development course, using a pretest–posttest cluster randomized trial design with one intervention group and one control group. Teachers served as the unit of randomization. Students, the primary unit of observation, 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 study was conducted from spring 2009 through spring 2010 (table 2.1). Outcomes were measured for teachers during both the 2008/09 and 2009/10 school years and for students during the 2009/10 school year. Teachers in the intervention group received a 24-hour Making Sense of SCIENCE™ professional development course on force and motion in summer 2009. They received no additional Making Sense of SCIENCE™ professional development or support during the school year.

Table 2.1. Experimental design and measurement points

	2009		*2009/10*	
Group	*Spring*	*Summer*	*Fall/Spring*	*Spring*
Teachers	ATLAST Test of Force and Motion for Teachers pretest; Teacher survey 1	Making Sense of SCIENCE™ professional development for intervention group	Teach force and motion	ATLAST Test of Force and Motion for Teachers posttest; Teacher survey 2
Students	State standardized tests in grade 7 mathematics (California Standards Test or Arizona Instrument to Measure Standards)		ATLAST Test of Force and Motion for Students before and after students receive instruction on force and motion	State standardized test in grade 8 science, physical science reporting clusters (California Standards Test only; no equivalent measure in Arizona)

Source: Author.

The counterfactual condition consisted of "business as usual." The control group teachers did not have access to the Making Sense of SCIENCE™ course during the study year. Like the intervention group teachers, they could participate in any other professional development that did not involve middle school force and motion. All control group teachers were offered the opportunity to take the Making Sense of SCIENCE™ course in the summer of 2010, after study data had been collected.

The intervention and control group teachers taught their lessons on force and motion in the first or second semester of the 2009/10 school year. Teachers made a commitment to take part in the study, but participating in the Making Sense of SCIENCE™ training and using what

they learned in the training in their classrooms were voluntary. In their classrooms, teachers used their usual local science curricula, textbooks, and other resources.

The timeline for gathering teacher measurements covered a calendar year, from administration and collection of pre-course outcome measures in spring 2009, before the professional development courses were run, to post-course measures in the winter/spring of 2009/10. Collection of data on students took place over two academic semesters. As part of the study, students took the ATLAST Test of Force and Motion in fall 2009 (pretest) and again within two weeks of their classroom unit on force and motion (posttest). Researchers obtained students' scores on state standardized achievement tests in the spring of 2009 (pretest) and a year later, in the spring of 2010 (posttest).

Site selection

Regional research sites were identified through discussions with district and county science educators in the western United States. Initial contacts were made through an extensive network of WestEd contacts; other contacts were identified in those conversations. Because of the large number of grade 8 science teachers needed for the study, the search for study schools focused on urban districts with at least 15 middle schools and larger geographic regions consisting of many districts with a smaller number of middle schools per district. The criteria for participation included the following:

- Stable district science program.
- Strong science leadership (as evidenced, for example, by a district staff position allocated to science curriculum coordination, an active cadre of science staff developers, or teacher leaders in science).
- No district or regional professional development in middle school force and motion within previous three years.
- No district or regional middle school science professional development initiatives involving case discussions or looking at student work within previous three years.
- Academically, culturally, and linguistically diverse student population.
- Proven ability to recruit teachers for professional development.
- Willingness to provide student test and demographic data from district administrative records.
- Availability of qualified professional educator willing to serve as local coordinator for the site.

The sites selected through this process included five in California (El Centro/Coachella, Pomona, Riverside/Lake Elsinore, San Diego, and San Joaquin) and one in Arizona (Tucson). Site coordinators were hired as consultants to oversee study activities at each site, including recruiting teachers, arranging for meeting and course facilities, running local meetings at which they collected teacher test and survey data, tracking down missing teacher or student

data as needed, and supporting local course facilitators and research staff with logistics as needed.

Depending on its size, each research site had one to three coordinators. Most coordinators were employed as science educators in county offices of education, school districts, or a local university. The group included three county or district science program coordinators and four science specialists teaching at the middle school level. Qualifications for serving as a site coordinator included extensive experience organizing and leading teacher professional development, strong local connections to teachers and district staff, and an orientation that was compatible with the Making Sense of SCIENCE™ professional development model, including a social constructivist perspective focusing on helping students and teachers learn about science through collaborative discourse.

Recruitment of teacher sample

Statistical power estimates (see appendix A) indicated that a teacher sample of 120 was needed to achieve 80 percent power to detect student impacts of 0.20 standard deviations or larger (0.23 standard deviations or larger for English language learners) and teacher impacts of 0.51 standard deviations or larger (for type I error = 0.05).

Coordinators at each of the six sites were asked to recruit a volunteer sample of up to 36 grade 8 science teachers, a recruitment target that exceeded the number needed, in order to allow for sample attrition. The number of teachers enrolled in the study from each district varied depending on teacher interest. Teachers were recruited by email and through announcements during professional meetings. They were considered eligible to participate if they were currently teaching grade 8 physical science in the 2008/09 school year, expected to be doing so in the 2009/10 school year, and had never taken a Making Sense of SCIENCE™ course. Teachers also had to consent to the study requirements, including the requirements to:

- Be randomly assigned to either the intervention or the control group.
- Attend two two-hour project meetings, one in winter/spring 2009 and one in winter/spring 2010.
- Attend a staff development course, Force and Motion for Teaching*,* in either summer 2009 (for intervention group teachers) or summer 2010 (for control group teachers).
- Teach and complete a classroom force and motion unit by March 31, 2010.
- Provide survey and test data for the course evaluation.

Participating teachers were volunteers and, thus, are not assumed to be representative of grade 8 science teachers in their schools, districts, or states.

Random assignment procedure

The study used teacher-level random assignment with school as a blocking factor when there were two or more teacher participants per school and a constructed stratum of two teachers as a blocking factor for teachers who were the only participants at their schools. A total of 181

teachers attended project baseline data collection meetings, after which they were randomly assigned to groups (90 to the intervention and 91 to the control group) (table 2.2).

Table 2.2. Number of teachers recruited and randomly assigned to intervention and control groups, by research site

Site	*Intervention group*	*Control group*	*Total*
1	14	15	29
2	13	14	27
3	15	15	30
4	15	15	30
5	18	18	36
6	15	14	29
All sites	90	91	181

Source: Author's analysis of primary data collected for the study.

For schools with two or more participating teachers, randomization was done within each school. All schools with only one teacher participant were ranked based on 2008 school-level state test scores.[3] The ranked list was then separated into blocks consisting of two teachers each. The first teacher in each block was randomly assigned to either the intervention or the control group and the second to the other group. This procedure was followed at each site.

The principal investigator of the study assigned teachers to groups. There were no breakdowns in random assignment or crossovers between groups. By the end of the study, however, some blocks had changed because of attrition, creating two additional situations: (a) singletons consisting of only one teacher because the other teachers were no longer in the study and (b) blocks that still had two teachers remaining but in which both teachers were now in the same condition. Appendix B provides details of assignment to blocks and procedures for resolving these situations.

Procedures to minimize contamination of control group teachers

One of the challenges of a design in which teachers are the unit of assignment within schools is that the close proximity of implementation and control group teachers increases the possibility of contamination of the control group. This is particularly true at the middle school level, where teachers typically work in subject area and grade-level teams that make detailed group decisions about curricula and instruction. In this study, there was a potential for control group teachers to learn about the content and approaches of the Making Sense of SCIENCE™ course and even to look at the materials from the course. Implementation group teachers could

[3] Schools in California were classified into blocks based on the 2008/09 school-level mean percentages of students scoring at or above proficient on the grade 8 California Standards Tests of mathematics and reading. Schools in Arizona were classified based on the 2008/09 school-level mean student scale scores on the grade 8 Arizona Instrument to Measure Standards in mathematics and reading.

also have spontaneously shared their newfound content knowledge or pedagogical strategies with their colleagues when they planned their force and motion lessons.

Several steps were taken to prevent crossovers between intervention and control groups. In project meetings in spring 2009 held at each site before teachers signed contracts to participate in the study, the regional site coordinator made a presentation to teachers on the threats of contamination. The aim was to enlist teachers' cooperation in maintaining the integrity of the random assignment by building an understanding of, and commitment to, the research process. At the meeting, the study team also asked all participants to sign both a consent agreement and a detailed Teacher Agreement to Protect the Study (see appendix C), both of which stipulated that they would preserve the differences between experimental and control groups by not sharing or receiving course materials or information for the duration of the study and that they would protect the validity of students' performance on tests by arranging for a proctor to administer the test, not helping students answer the questions, and not looking at or copying the test questions.

Teachers' post-instruction survey responses (see appendix D) indicate that despite these procedures to protect the integrity of random assignments, there may have been some contamination. Four intervention group teachers and four control group teachers indicated that one or two teachers in their school who did not participate in the Making Sense of SCIENCE™ course had implemented aspects of the course.

Although worth considering, these responses are not of serious concern for several reasons. First, the survey question about implementing "any aspects" of the course was vague; it is possible that the teachers were referring to aspects that the two groups' instruction shared, rather than actual contamination between the groups. Second, the number of teachers expressing these concerns was small (about 6 percent of participating teachers). Third, if contamination occurred, it would mean that the true effects of the intervention were actually larger than those measured, not that the results were discredited.

Parent consent procedures

The Institutional Review Board at Independent Review Consulting, Inc.[4] required active parental consent to collect ATLAST and student standardized test score data. Many of the school districts participating in the study also required active parental consent before releasing state test score data.[5] Parental consent was solicited through a letter and consent form that was sent home with each student (see appendix E). The consent form described the purpose of the research and detailed the data for which the study team was requesting consent.

[4] Independent Review Consulting, Inc. (IRC: irb-irc.com) is a fully accredited IRB review service that fulfills the role of an institution as defined in the Common Rule, and FDA regulations. This institution provides IRB services for research regulated by other agencies.

[5] Because the research team was barred from collecting student background information or test score information from students whose parents did not provide consent, it was not possible to compare the characteristics of participating and nonparticipating students or differences in characteristics between participating and nonparticipating students across the intervention and control groups.

Data collection instruments

Outcomes were measured for intervention and control group teachers and students through data collected during both the 2008/09 and 2009/10 school years (table 2.3). Teacher pre- and post-course surveys and tests were administered in the spring before and the winter after the professional development courses, which occurred in summer 2009. Students were given a science content pretest and a posttest within two weeks before and two weeks after their classroom instruction on force and motion. Students' scores were obtained for standardized achievement tests at the end of the academic year preceding and the year in which the experiment was conducted. Video recordings of all professional development course sessions and detailed attendance records were collected to allow analysis of fidelity of implementation.

The intervention evaluated in this study is a teacher course designed to strengthen teachers' science and pedagogical knowledge in a way that is compatible with whatever student curriculum is already used in the classroom. The intervention is not a student curriculum. No materials were provided for use in teachers' classrooms, although some teachers did adapt activities they completed in the course for student use. Classroom observation data were to have been collected in a small sample of participating teachers' classrooms, but resource constraints prevented the collection of most of those data.

Table 2.3. Measurement instruments, samples, schedule, and data collection procedures, by data collection instrument

Instrument	*Variable measured*	*Sample*	*Procedure*
Student measure			
Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion for Students (pretest and posttest)	Knowledge of force and motion	Physical science students in two randomly selected grade 8 classes of each teacher participating in study (n = 5,130)	Proctors administered tests before the force and motion was taught and within two weeks after it was taught
2009/10 Grade 8 California Standards Test physical science reporting clusters	Knowledge of physical science	California physical science students in two randomly selected grade 8 classes of each teacher participating in study (n = 3,771)	Obtained from district administrative records
2008/09 Grade 7 mathematics (California Standards Test or Arizona's Instrument to Measure Standards)	Entering academic performance level	Physical science students in two randomly selected grade 8 classes of each teacher participating in study (n = 4,454)	Obtained from district administrative records
Student and school information survey	Student population, curricular and school context information	All classes in which student data were collected (n = 249 classes)	Teachers completed at time of student posttest (fall/winter 2009/10)
Teacher measure			
ATLAST Test of Force and Motion for Teachers (pretest and posttest)	Knowledge of force and motion	All participating teachers (n = 133)	Site coordinators administered to teachers at meetings in winter/spring 2008/09 and one year later
Teacher survey 1 (baseline) and teacher survey 2 (postinstruction)	Teacher background, beliefs, and practices related to teaching force and motion	All participating teachers (n = 133)	Site coordinators administered to teachers at meetings in winter/spring 2008/09 and one year later
Course implementation			
Video recordings of professional development sessions	Fidelity of implementation	All course sessions at each research site (n = 30)	Course facilitator video recorded all sessions
Attendance records	Intervention dosage	All teachers in intervention group (n = 69)	Facilitator recorded arrival and departure times of each participant for each course session

Source: Author.

Student measures

Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion for Students. Students' science content knowledge was measured using a test that was developed and validated as part of the ATLAST project, by Horizon Research, Inc., in collaboration with Project 2061 of the American Association for the Advancement of Science (Smith and Banilower 2006a, 2006b). ATLAST was funded by the National Science Foundation to provide rigorous and well-validated measurement instruments to be used in evaluations of science education programs. In this study, the study team used the ATLAST Test of Force and Motion for Students (http://www.horizon-research.com/atlast/). This multiple-choice test

measures science content in the National Science Education Standards and reflects the research literature documenting misconceptions related to science concepts in these domains. The test, administered in one 45-minute period, is composed of 27 multiple-choice items. Scores are computed as the percentage of questions answered correctly. The test has an alpha reliability coefficient of 0.86; the alpha coefficient of the student test based on data collected in this study was 0.82.

Grade 8 California Standards Test physical science reporting clusters. Student scores on the 2009/10 Grade 8 California Standards Test in science were obtained from districts' administrative records for use as an outcome variable in the student-level analyses.[6] Physical science scores were available in two reporting clusters—motion (8 items) and forces, density, and buoyancy (13 items). The two clusters are designed to measure 17 California science content standards (see appendix F). Analyses were conducted based on the percentage of the 21 items in these two reporting clusters that were answered correctly.

As with other state tests, all questions on the California Standards Test are evaluated by committees of content experts, including teachers and administrators, to ensure the questions' appropriateness for measuring the state academic content standards in middle school science. In addition to being reviewed for content, all items are reviewed and approved to ensure their adherence to principles of fairness and to ensure that no bias exists with respect to characteristics such as gender, race/ethnicity, or language. Reported reliability figures for the test in science range from 0.88 to 0.91.

Grade 7 standardized mathematics test. The 2008/09 California Standards Test and the 2008/09 Arizona Instrument for Measuring Standards (AIMS) grade 7 mathematics scores were obtained from district administrative records. In both student and teacher impact analysis models, scaled grade 7 student scores for mathematics from 2008/09 were used as a covariate measure of student entering academic performance level.

Student data from administrative records. Student demographic information (see appendix G) was obtained from district administrative records. Variables collected included race/ethnicity, sex, and English language learner classification. Institute of Education Sciences guidelines were followed with regard to reporting race/ethnicity in the categories of White, Black, Hispanic, Asian, American Indian, Other, and multiple race/ethnicity. For English language learners, scaled scores on the state-administered California English Language Development Test (CELDT) or the Arizona English Language Learner Assessment (AZELLA) were collected. Districts were asked to report each student's English language learner classification as of the beginning of the 2009/10 school year, in the following categories: English Only, Initially Fluent English Proficient (nonnative English speakers classified as fluent in English when they arrived in the district), English Language Learner; and Reclassified Fluent English Proficient (English Language Learners who were reclassified as fluent in English after some time in the district). Districts and states differ in their criteria for classifying a student as an

[6] These scores were collected in California only, because physical science scores are not reported separately from total science scores on Arizona's grade 8 Arizona Instrument for Measuring Standards (AIMS) test. Statistical power was judged adequate for estimating program impacts on student outcomes using the California subsample. See study power estimates in appendix A for more details.

English language learner or as fluent in English; in the analyses reported here, the district's classification defined the variable.

Student and school information survey. For each class in which student data were collected, teachers were asked to complete a classroom information survey that included questions on the number of students in the class in each of several categories, including special education students, students eligible to receive free or reduced-price meals, gifted students, and so forth; the school locale (urban, rural, and so forth); and the science curriculum used in the class.

Teacher measures

Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion for Teachers.[7] Teachers' science content knowledge was measured using the ATLAST Test of Force and Motion for Teachers (http://www.horizon-research.com/atlast/). The test, with a reported reliability of 0.84, includes 25 multiple-choice items that measure teachers' science content knowledge, ability to use it to diagnose student thinking, and ability to use it to make instructional decisions (Smith and Banilower 2006b). Scores are computed as the percentage of questions answered correctly. The alpha coefficient of the teacher test based on data collected in this study was 0.82.

Teacher surveys. All participating teachers were asked to complete a pre-course survey in spring 2009, preceding the intervention, and a post-instruction survey the following year, after they had taught their classroom units on force and motion. These surveys had been used in numerous studies over the past 10 years to measure teachers' self-reported outcomes of Making Sense of SCIENCE™ courses with content-specific survey questions changed for studies in different science domains (Heller, Daehler, and Shinohara 2003, 2011; Heller and Kaskowitz 2004; Heller, Shinohara, Miratrix, Rabe-Hesketh, and Daehler 2010). Because self-report data are of limited use in judging course impacts, survey results were intended for descriptive purposes only and not as the basis for inferences about efficacy.

The survey development process began in July 1999 when Heller Research Associates conducted a search for teacher surveys measuring impact of science professional development. None of the available instruments was sufficiently well-aligned with the Making Sense of SCIENCE™ professional development model and intended outcomes. As a result, research and program staff collaborated to identify constructs and kinds of information required in six domains: teachers' educational background and science teaching experience, classroom instructional practices, beliefs about science teaching and children's learning, confidence in their ability to teach force and motion, and self-reported impact of courses on teaching.

It was important in the development process to be sure that the type of information that the survey would yield would be useful and relevant for multiple audiences: course developers, teachers, policy makers, and the educational research community. Therefore, focus groups were conducted with teachers and with program developers for the purpose of identifying which aspects of the teachers' backgrounds, experiences, and outcomes were most important

[7] The assessment was developed by the Assessing Teacher Learning About Science Teaching (ATLAST) project at Horizon Research, Inc. ATLAST is funded by the National Science Foundation under grant number EHR-0335328.

to them. The team conducted a total of four focus groups from the fall of 1999 to spring of 2000.

Program and research staff then drafted survey questions in each domain. After careful internal review and editing, draft pre-course and post-instruction instruments, originally containing 65 and 50 questions respectively, were tested to identify problems with navigation and comprehension in a series of cognitive interviews. The surveys were administered individually to a sample of teachers drawn from the population to be surveyed to determine whether teachers interpreted the items as intended or misunderstood anything about the items. Six interviews were conducted in the first round of cognitive testing. Subsequently, the instrument was revised to address identified problems and then the revised instrument was tested with 10 respondents.

The surveys were then used in pilot and national field test studies in which they were administered before and after teachers completed Making Sense of SCIENCE™ professional development courses from March 2000 through December 2005. The current study was the first to use data from these surveys to test a specific teacher outcome: confidence in ability to teach force and motion. Twenty-three of the survey items were selected to measure teacher confidence (see appendix H), including:

- Confidence in their ability to teach force and motion content that appears in state curriculum standards (nine items, for example, "An object that is moving with constant speed can have changing velocity").
- Confidence about implementing general teaching goals and strategies (nine items, for example, "Teach students to collect and carefully record data").
- Overall confidence in teaching (five items, for example, "I know how to use the district force and motion curriculum").

The generality of the items makes them appropriate indicators of teacher confidence for any grade 8 physical science teacher, not just teachers exposed to the intervention.

Before using them in this study, the reliability of these 23 items was computed based on data collected in an earlier randomized study (Heller et al. 2010), yielding an alpha coefficient of 0.86. Based on data collected in the current study, the reliability was 0.90. One overall confidence index was computed for each teacher based on his or her individual ratings on the 23 survey items. Each teacher's overall confidence score was computed as the mean of the confidence ratings provided by that teacher. Because the overall measure is based only on the items to which each teacher responded, no correction for missing items was needed.

Documentation of course implementation

Facilitators video recorded all 30 course sessions at each research site. Instructions for collecting these recordings were distributed to facilitators in a protocol appended to the course facilitator guide (see appendix I).

Facilitators also recorded the arrival and departure times of each participant in each course session, using an attendance recording form provided by the study team (see appendix J). They also recorded the actual length of each session. These records documented the amount of exposure each participant had to the intervention compared with the number of hours a participant could potentially have had.

Data collection procedures

Administration of student tests

Before student test administration, packets of student tests were sent to participating teachers. These packets included instructions and an administration script (see appendix K), as well as a classroom information survey to be completed by the teacher about the class. Each student testing package included instructions on how to administer the tests, including rules on opening, distributing, and collecting the tests; securing completed answer sheets in sealed envelopes; and returning them for data processing and scoring. Arrangements were made at each school for a professional staff person who was not directly involved in that classroom (for example, a counselor, aide, administrator, or other teacher) to administer the science tests, following a detailed testing protocol provided by the research team. These test proctors were often colleagues of teachers involved in the study, whose participation was consistent with common practice for the administration of standardized tests in schools. Proctors completed an honorarium request form, specifying the teacher and class sections in which they administered each test, to verify their participation.

Proctors administered the pretests during one period of class time in fall/winter 2009. They administered posttests within two weeks after the completion of the class's instruction in force and motion, whenever that occurred during the school year (see table 2.1). Students who missed a test because they were absent were given a make-up test as soon as they returned to school. Instructional lessons on force and motion took place over four to eight weeks, and no efforts beyond teachers' usual practices were made to provide make-up instruction for students who were absent during any lessons.

The data process team applied quality assurance procedures to verify that the student data they received were accurate and secure. These procedures included matching names, checking test forms, comparing student identification numbers and dates of birth on pretests and posttests, and verifying parental consent for each student.

Administration of teacher tests and surveys

Site coordinators administered science content tests and surveys to both intervention and control group teachers at regional project meetings in winter/spring 2009, before random assignment to the treatment or control group, and again in fall/winter 2010, after teachers had completed teaching the unit on force and motion, and students had taken their posttests. Site coordinators were provided with detailed test administration instructions (see appendix L) to standardize procedures across research sites.

Collection of course implementation data

Facilitators returned the videotapes and attendance sheets to the research staff at the end of the course. Course facilitators recorded the attendance of each participant in each course session, as well as the length of each session. These records documented the amount of exposure each participant had to the intervention.

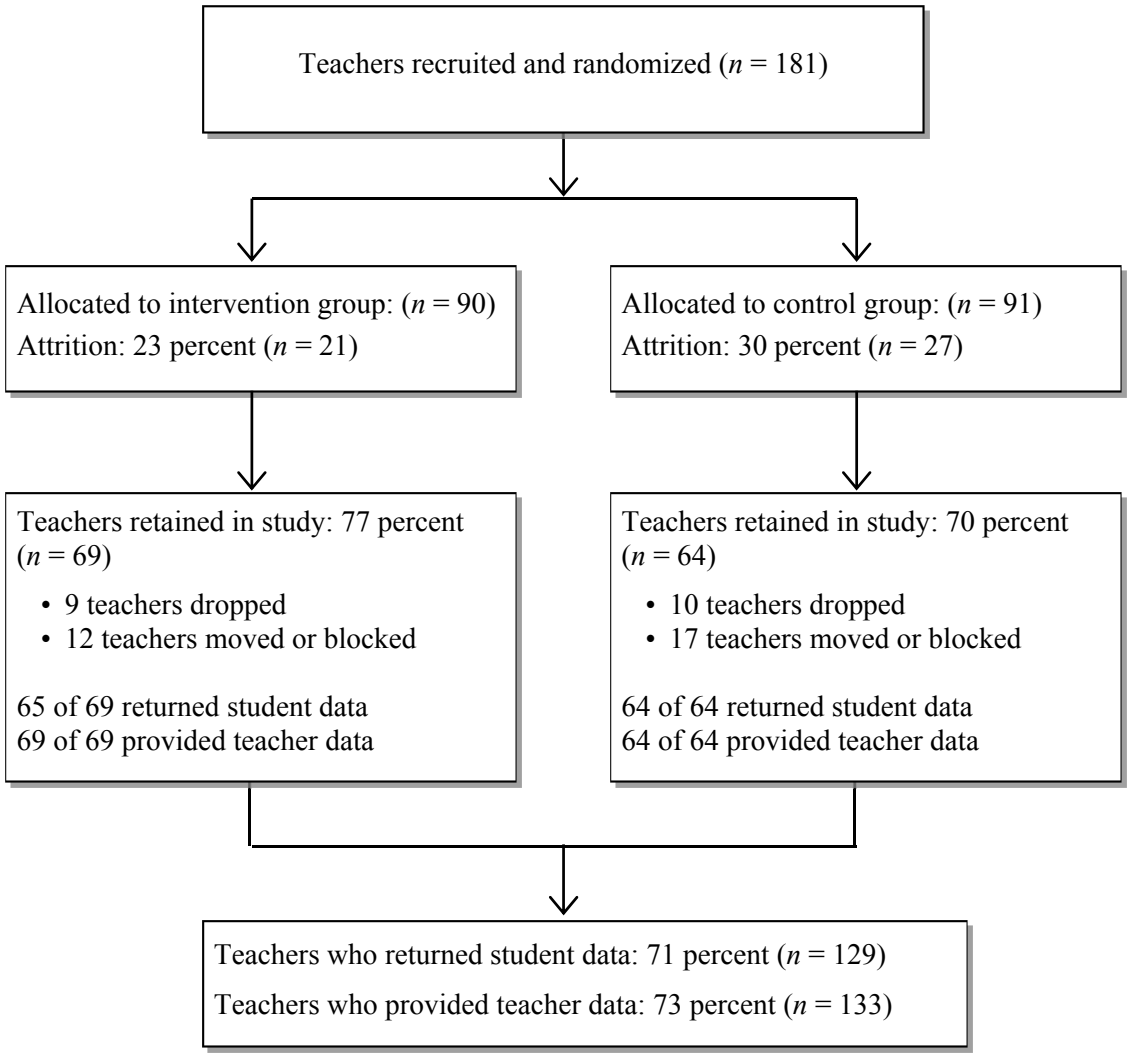
Teacher analytic sample

The teacher analytic sample was defined and tracked based on the Consolidated Standards of Reporting Trials (CONSORT) to document the flow of participants through each stage of the randomized trial (figure 2.1). The confirmatory intermediate-level analytic sample included all teachers randomly assigned to treatment or control condition for whom valid posttest data were available.

Of the 181 teachers originally recruited and randomly assigned, 73 percent completed the study and provided teacher survey data (77 percent of intervention group teachers and 70 percent of control group teachers) and 71 percent of the teachers provided student posttest data (72 percent of intervention group teachers and 70 percent of control group teachers) (see figure 2.1). Nine intervention group teachers (10 percent) and 10 control group teachers (11 percent) dropped out; the rest were not retained for reasons outside of their control. Teachers who left the study were categorized as *dropped* if they left for personal reasons, *moved* if they were no longer teaching in eligible classes in a study research site, or *blocked* if their district or school did not approve their participation.

During the study period, both California and Arizona made severe budget cuts. As a result, 12 percent of intervention group teachers and 18 percent of control group teachers lost their teaching positions or had to change grade levels. The proportion of each group retained varied considerably across research sites (table 2.4).

Figure 2.1 Consolidated Standards of Reporting Trials (CONSORT) diagram for teachers providing data



Note: Categories of attrition are *dropped,* if a teacher left for personal reasons (for example, pregnancy or illness) or because of time conflicts; *moved*, if a teacher left the teaching profession, was laid off, transferred to a nonparticipating district, or ended up not teaching grade 8 science; and *blocked*, if a teacher taught in a district or school that did not approve participation in the study.
Source: Author's analysis of primary data collected for the study.

Table 2.4. Number of teachers recruited and retained, by site and experimental condition

Sample	*Number recruited and randomly assigned*	*Number providing teacher data*	*Number providing student data*	*Percent retained in study*[a]
Full sample	181	133	129	73.5
Intervention group	90	69	65	76.7
Control group	91	64	64	70.3
Site 1				
Intervention group	14	12	12	85.7
Control group	15	11	11	73.3
Site 2				
Intervention group	13	8	7	61.5
Control group	14	9	9	64.3
Site 3				
Intervention group	15	10	10	66.7
Control group	15	14	14	93.3
Site 4				
Intervention group	15	11	8	73.3
Control group	15	8	8	53.3
Site 5				
Intervention group	18	14	14	77.8
Control group	18	11	11	61.1
Site 6				
Intervention group	15	14	14	93.3
Control group	14	11	11	78.6

Source: Author's analysis of primary data collected for the study.
[a] Number of teachers providing teacher survey data divided by number of teachers randomly assigned.

Baseline equivalence of intervention and control group teacher samples

The internal validity of the study depends on baseline equivalence between intervention and control group teachers. Teacher-level characteristics were compared for the teacher samples that were randomly assigned to the intervention and control groups (recruited) that remained in the sample through the conclusion of the study (retained) and those that left the study before its conclusion (not retained).

Baseline science content test scores of intervention group teachers were more than 0.25 standard deviation higher than scores of control group teachers, as measured by the ATLAST Test of Force and Motion for Teachers (table 2.5), but the differences were not

significant for either the full recruited sample or the retained sample. Nevertheless, the magnitude of the differences prompted a review of the random assignment procedures, which the study team confirmed had been carried out correctly. The differences between groups were statistically controlled for in the impact analyses by including teacher pretests as a covariate in both the teacher and student models.

There were no statistically significant baseline differences between the treatment and control group teachers in any of the study samples (the sample recruited, that retained through follow-up, or that not retained) for the measure of teacher confidence in ability to teach force and motion.

We also found no statistically significant differences in the demographic characteristics of intervention and control teachers in any of the three sample subgroups (see appendix M). For example, among both treatment and control group teachers retained in the sample, about 60 percent were women, 73 percent were White, and 87 percent were native English speaking.

The only comparison for which a significant difference between intervention and control groups was detected was in the number of semesters of postsecondary classes taken in science in the retained teacher sample: control group teachers took more such classes than intervention group teachers (see appendix M). Participants were generally experienced teachers, with all samples averaging about 11 years of teaching experience, 9 years of experience teaching science, 6 years of experience teaching force and motion, and more than 8 years of experience teaching English language learners.

Table 2.5. Teacher baseline measures on outcome variables for teacher sample recruited, retained, and not retained, by experimental condition

Measure	*Intervention group*	*Control group*	*Difference*	p-*value*[a]
Teacher pretest of force and motion				
Full teacher sample				
Mean percent correct	55.8	51.1	4.8	.05
Standard deviation	17.2	16.7		
n	90	91		
Teacher sample retained				
Mean percent correct	57.4	52.1	5.3	.08
Standard deviation	17.2	17.0		
n	69	64		
Teacher sample not retained				
Mean percent correct	51.2	48.7	2.5	.66
Standard deviation	16.6	15.9		
n	21	27		
Confidence in ability to teach force and motion				
Full teacher sample				
Mean	2.4	2.4	0	.99
Standard deviation	0.5	0.5		
n	89	91		
Teacher sample retained				
Mean	2.4	2.5	–0.1	.41
Standard deviation	0.4	0.4		
n	68	64		
Teacher sample not retained				
Mean	2.4	2.4	0	.84
Standard deviation	0.6	0.5		
n	21	27		

a. Two-tailed Fisher's exact test for equality of proportion between intervention and control group teachers.
Source: Author's analysis of primary data collected for the study.

Student analytic sample

The student sample was identified at the class level through random selection of two grade 8 physical science classes per retained teacher. All physical science classes were considered eligible except those that included only special education students. The classes were determined using a class selection worksheet (see appendix N) that led teachers through a procedure for selecting classes using a table based on random numbers.

Of the 133 retained teachers, almost all submitted student data in 249 class sets (127 from intervention group classes and 122 from control group classes) (table 2.6). The numbers of intervention group and control group teachers who submitted two class sets were identical (60 each).

Table 2.6. Number of class sets submitted, by experimental condition and site

Sample/site	*Number of teachers submitting one class set*	*Number of teachers submitting two class sets*	*Number of class sets submitted*
Intervention group	7	60	127
Control group	2	60	122
Full sample	9	120	249
Site 1			
Intervention group	#	#	23
Control group	#	#	21
Site 2			
Intervention group	0	8	16
Control group	0	8	16
Site 3			
Intervention group	#	#	19
Control group	0	13	26
Site 4			
Intervention group	#	#	18
Control group	0	7	14
Site 5			
Intervention group	#	#	27
Control group	#	#	23
Site 6			
Intervention group	#	#	24
Control group	0	11	22

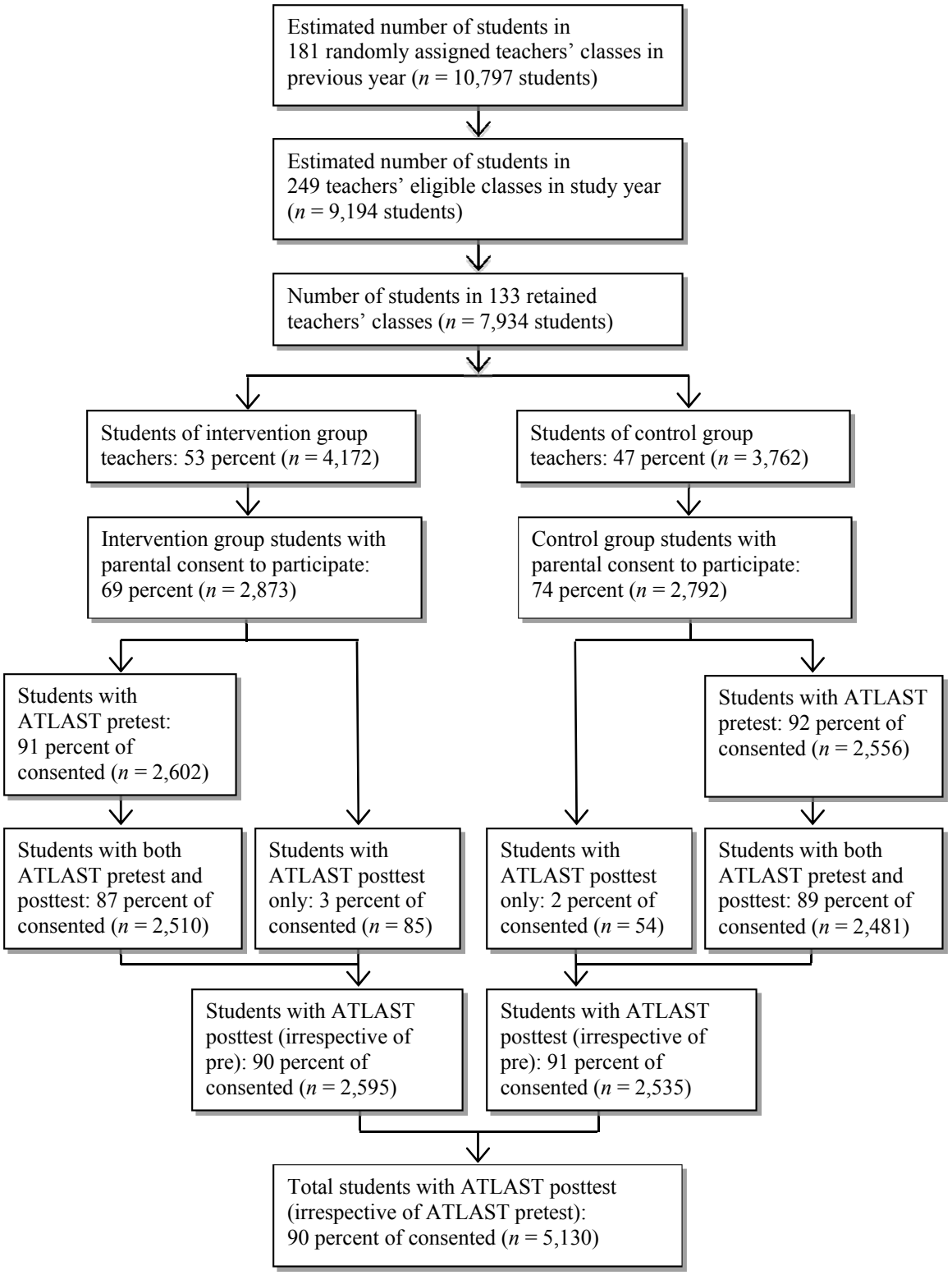
Note: # indicates values were suppressed to reduce disclosure risk.
Source: Author's analysis of primary data collected for the study.

Of the 181 grade 8 science teachers randomly assigned during the school year preceding the student data collection year, 27 could not participate in the study because they were no longer teaching the target student population during the study year (see figure 2.1). This left 154 teachers whose classes might have been eligible to participate in the study. About 9,200 students would have been eligible to participate in the study if all of their teachers had been retained (figure 2.2). Of the 154 eligible teachers, 133 (86 percent) were retained in the study, 131 of whom provided student data. Information was not available on students of teachers who left the study before student data were collected. Thus, the data do not represent all students who were eligible for the study.

Two overlapping analytic samples were used: all students in participating teachers' classrooms that were eligible for this study (that is, classes that did not consist exclusively of special education students) and English language learner students in those classrooms. The analytic samples included all students whose parents provided consent and who had valid ATLAST posttest score data.

Consent rates were calculated by dividing the number of students for whom parental consent was granted by the number of students enrolled in the classrooms for which teachers provided student data. Consent rates were 69 percent for students of intervention group teachers and 74 percent for students of control group teachers. Response rates on the ATLAST Test of Force and Motion were calculated by dividing the total number of students with valid posttest data by the number of students with parental consent to participate in the study. Response rates were 92 percent for students of intervention group teachers and 93 percent for students of control group teachers.

Figure 2.2 Consolidated Standards of Reporting Trials (CONSORT) diagram for students providing data



Source: Author's analysis of primary data collected for the study.

Baseline equivalence of intervention and control group student samples

Students' baseline scores on the key outcome variables were analyzed for equivalence across the two study groups (table 2.7). No statistically significant differences were detected between intervention and control group students on pretest science content, as measured by the ATLAST Test of Force and Motion, state standardized test scores in mathematics from the previous school year, or state tests of English language development.

Table 2.7. Teacher-level means on key student measures at baseline, by experimental condition

Measure	*Intervention group*	*Control group*	*Difference*	*p*[a]
ATLAST Test of Force and Motion pretest (full student sample)				
Mean percent correct	36.6	36.6	0	.99
Standard deviation	5.8	5.9		
N of teachers	67	60		
N of students	2,611	2,540		
2008/09 Grade 7 mathematics (California Standards Test or Arizona's Instrument to Measure Standards) (full student sample)				
Mean scale score	402.9	402.7	0.2	.99
Standard deviation	85.8	83.0		
N of teachers	63	61		
N of students	2,258	2,345		
ATLAST Test of Force and Motion pretest (English language learner sample)				
Mean percent correct	30.7	30.0	0.7	.67
Standard deviation	7.8	8.1		
N of teachers	46	46		
N of students	198	257		
2008/09 Grade 7 mathematics (California Standards Test or Arizona's Instrument to Measure Standards) (English language learner sample)				
Mean scale score	344.1	328.6	15.5	.29
Standard deviation	71.0	67.5		
N of teachers	47	45		
N of students	186	231		
Fall 2009 test of English language development, overall proficiency (English language learner sample)				
Mean scale score	577.7	589.3	–11.6	.30
Standard deviation	50.1	53.3		

Measure	*Intervention group*	*Control group*	*Difference*	*p*[a]
N of teachers	43	44		
N of students	206	246		
Fall 2009 test of English language development, speaking proficiency (English language learner sample)				
Mean scale score	569.1	589.0	–19.9	.20
Standard deviation	71.6	73.9		
N of teachers	44	47		
N of students	208	258		
Fall 2009 test of English language development, listening proficiency (English language learner sample)				
Mean scale score	594.1	603.6	–9.5	.45
Standard deviation	53.9	62.6		
N of teachers	42	45		
N of students	203	252		

a. *F*-test from ANOVA was used to test whether the student measures at baseline in the intervention and control groups were equivalent.
Source: Author's analysis of primary data collected for the study.

There also were no significant differences between the intervention and control groups in their demographic characteristics (table 2.8). The student sample included slightly more girls (52 percent) than boys (48 percent). In terms of racial/ethnic composition, both groups included about 44 percent Hispanic, 30 percent White, 10 percent Asian, and 7 percent Black students. On average, about 10 percent of the students in both the intervention and control groups were English language learners.

Table 2.8. Student demographic information, by experimental condition

	Intervention group		*Control group*		
Characteristic	*Number*	*Percent*[a]	*Number*	*Percent*[a]	*p*-value[b]
Sex					.23
Female	1,355	52.5	1,391	51.6	
Male	1,222	47.4	1,300	48.2	
Unknown	#	#	7	0.3	
Ethnicity/race					.44
Hispanic	1,119	43.4	1,210	44.9	
White	839	32.5	831	30.8	
Asian	279	10.8	276	10.2	
Black	170	6.6	197	7.3	
Pacific Islander	74	2.9	88	3.3	
American Indian	41	1.6	30	1.1	
More than one race	34	1.3	36	1.3	
Other	11	0.4	10	0.4	
Unknown	12	0.5	20	0.7	
English language learner classification					.18
Native English speaker	1,587	61.5	1,581	58.6	
Initially fluent English proficient	206	8.0	211	7.8	
Reclassified fluent English proficient	507	19.7	573	21.2	
English language learner	223	8.7	269	10.0	
Unknown	56	2.2	64	2.4	

Note: White includes European; Black includes African American; Hispanic includes Latino and other Spanish origin; Asian includes Chinese, Indian, Japanese, Korean, and Vietnamese; Pacific Islander includes Filipino, Guamanian or Chamorro, Native Hawaiian, Samoan, and other Pacific Islander; American Indian includes Alaska Native.
# indicates values were suppressed to reduce disclosure risk.
a. Components may not sum to 100 because of rounding.
b. Chi-square test for equality of proportion between intervention and control students.
Source: Author's analysis of primary data collected for the study.

Districts and schools represented in sample

The 181 teachers came from 137 schools in 55 school districts (table 2.9). Research sites of the recruited teacher sample included between 4 and13 districts, each with between 19 and 30 schools. The 133 teachers retained in the analytic sample came from 102 schools in more than 40 districts (table 2.10). After attrition, the research sites included up to 10 districts with between 13 and 21 schools per district.

Table 2.9. Numbers of teachers, districts, and schools represented in recruited sample, by research site

Site	*State*	*Number of teachers*	*Number of school districts*	*Number of schools*
1	California	29	11	19
2	California	27	13	22
3	California	30	8	21
4	California	30	4	22
5	California	36	10	30
6	Arizona	29	9	23
Full sample		181	55	137

Source: Author's analysis of primary data collected for the study.

Table 2.10. Numbers of teachers, districts, and schools represented by retained teachers, by research site

Site	*State*	*Number of retained teachers*	*Number of school districts*	*Number of schools*
1	California	23	10	16
2	California	17	9	13
3	California	24	8	17
4	California	19	#	14
5	California	25	10	21
6	Arizona	25	9	21
Full sample		133	#	102

# Indicates values were suppressed to reduce disclosure risk.
Source: Author's analysis of primary data collected for the study.

Because teachers were drawn from many districts and schools, the number of study participants from any one district or school was generally small (tables 2.11 and 2.12). Fifty-four percent of the teachers taught in districts with four or fewer study participants, and approximately 73 percent were from districts with fewer than eight participants. About 57 percent of the retained teachers (76 of 133) were the only study participants from their schools (61 percent of intervention group teachers and 53 percent of control group teachers) (table 2.12). The remaining teachers were at schools with two, or at most three, teachers in the study.

Table 2.11. Numbers of retained teachers per district, by experimental condition

Number of teachers per district	*Number of districts*	*Intervention group*		*Control group*		*Full sample*	
		Number of teachers	*Percentage of teachers*	*Number of teachers*	*Percentage of teachers*	*Number of teachers*	*Percentage of teachers*
1	22	14	20.3	8	12.5	22	16.5
2	10	10	14.5	10	15.6	20	15.0
3	6	9	13.0	9	14.1	18	13.5
4		4	5.8	8	12.5	12	9.0
5-7	4	12	17.4	13	20.3	25	18.9
8 or more	3	20	29.0	16	25.0	36	27.1
Total	#	69	100.0	64	100.0	133	100.0

# indicates values were suppressed to reduce disclosure risk.
Source: Author's analysis of primary data collected for the study.

Table 2.12. Numbers of retained teachers per school, by experimental condition

Number of teachers per school	*Intervention group*		*Control group*		*Full sample*	
	Number	*Percent*	*Number*	*Percent*	*Number*	*Percent*
1	42	60.9	34	53.1	76	57.1
2	20	29.0	22	34.4	42	31.6
3	7	10.1	8	12.5	15	11.3
Total	69	100.0	64	100.0	133	100.0

Source: Author's analysis of primary data collected for the study.

School characteristics

The study team examined the characteristics of schools of teachers who were randomly assigned to experimental conditions (*recruited*), schools that had one or more teachers who remained in the study until its conclusion (*any retained*), and schools that had no teachers who remained in the study (*none retained*) (table 2.13). Enrollments in schools of recruited teachers as well as of teachers who remained in the study averaged about 900 students. About 20 percent of the students served by the schools of recruited teachers were English language learners, and more than half of all students were eligible for free or reduced-price meals. There were statistically significant differences in characteristics of the student populations served by schools with and without teachers who remained in the study. Schools of teachers who left the study were more urban and had higher proportions of English language learners. Although no statistically significant differences were found between the overall academic performance indexes of schools with and without retained teachers, average grade 8 standardized test scores in science were lower in schools whose teachers did not remain in the study.

Table 2.13. School-level characteristics of teacher sample, by retention status of teachers

Characteristic	*Schools with recruited teachers*	*Schools with any retained teachers*[a]	*Schools with no retained teachers*[a]	*Difference*	*p*[b]
Number of schools	137	102	35		
Setting (percent)					
Urban	37.7	34.3	47.2	–12.9*	.02
Suburban	45.7	45.1	47.2	–2.1	
Rural	13.8	16.7	<10	11.1	
Other/Unknown	2.9	3.9	0	3.9	
n	137	102	36		
Enrollment (number of students)					
Mean	889.2	890.4	885.6	4.8	.55
Standard deviation	390.9	341.0	512.4		
Range	12–3,055	12–1,727	108–3,055		
n	137	102	36		
English language learners (percent)					
Mean	20.0	17.8	26.5	–8.7**	<.01
Standard deviation	15.2	14.0	16.7		
n	136	101	35		
Eligible for free or reduced-price meals (percent)					
Mean	56.5	54.0	63.5	–9.5	.07
Standard deviation	26.8	26.6	26.5		
n	135	100	35		
Academic performance index (API)					
Mean	753.3	761.7	733.9	27.8	.15
Standard deviation	80.8	77.3	86.6		
n	112	78	34		
Academic performance index for English language learners					
Mean	693.0	696.0	685.6	10.4	.46
Standard deviation	58.8	60.8	54.0		
n	102	72	30		
Percent at or above proficient on 2009 grade 8 standardized test in science					
Mean	51.5	54.8	42.2	12.6**	<.01
Standard deviation	21.1	19.9	22.1		
n	137	102	36		

*Significantly different from zero at the 0.05 level, two-tailed test. **Significantly different from zero at the 0.01 level, two-tailed test.
a. Components may not sum to 100 because of rounding.
b. Exact Wilcoxon rank sum test between schools with any retained teachers and no retained teachers.
Source: Author's analysis of primary data collected for the study.

The percentage of schools with only intervention group teachers (81 percent) was higher than the percentage of schools with only control group teachers (69 percent) or schools with both groups of teachers (70 percent) (table 2.14). After attrition the sample of schools with only control group teachers (n = 36) was smaller than that of schools with only intervention group teachers (n = 43). It included fewer English language learners and students eligible for free or reduced-price meals. The academic performance index (API) of schools with only intervention group or only control group teachers was about 750, the API of English language learners at those schools was about 700, and 15–17 percent of grade 8 students at those schools scored at or above proficient on a standardized science test.

Table 2.14. School-level characteristics for retained teacher sample, by experimental condition

Characteristic	*Intervention group only*	*Control group only*	*Full sample*	*Total*
Number of schools	43	36	23	102
Percentage of schools with any retained teachers	81.1	69.2	69.7	73.9
Locale (percent)				
Urban	30.2	38.9	34.8	34.3
Suburban	39.5	47.2	52.2	45.1
Rural	25.6	11.1	8.7	16.7
Other/ Unknown	4.7	2.8	4.3	3.9
n	43	36	23	102
Enrollment (number of students)				
Mean	771.0	910.4	1,082.5	890.4
Standard deviation	336.1	320.1	296.9	341.0
Range	12–1,727	388–1,493	464–1,682	12–1,727
n	43	36	23	102
English language learners (percent)				
Mean	6.7	5.5	8.6	6.7
Standard deviation	13.5	12.7	14.7	13.4
n	43	36	22	101
Eligible for free or reduced-price meals (percent)				
Mean	16.3	14.3	22.0	16.8
Standard deviation	30.4	28.2	32.9	30.1
n	43	35	22	100
Academic performance index				
Mean	746.8	774.5	768.3	761.7
Standard deviation	75.0	86.0	65.4	77.3
n	32	28	18	78

Characteristic	*Intervention group only*	*Control group only*	*Full sample*	*Total*
Academic performance index for English language learners				
Mean	686.4	705.7	696.6	696.0
Standard deviation	61.4	64.4	55.5	60.8
N	28	26	18	72
Percent at or above proficient on 2009 grade 8 standardized test in science				
Mean	16.5	15.6	21.8	17.3
Standard deviation	27.8	27.9	30.2	28.2
N	43	36	22	101

Source: Author's analysis of primary data collected for the study.

The characteristics of the classes in which student data were collected did not differ significantly between intervention and control groups (table 2.15). In both groups class sizes averaged just under 30, and 8 percent of students were classified as special education students, 12 percent as gifted or honors, and almost half as eligible for free or reduced-price meals.

Table 2.15. Characteristics of classes that provided student data, by experimental condition

Measure	*Intervention*	*Control*	*Difference*	p[a]
Class size (number of students)				
Mean	29.3	29.3	–0.01	.94
Standard deviation	6.4	5.7		
n	126	114		
Physical science student population (percent)				
Special education or resource				
Mean	8.0	7.9	0.1	.85
Standard deviation	10.4	13.2		
n	125	113		
Gifted or honors				
Mean	12.5	12.0	0.5	.46
Standard deviation	21.9	21.5		
n	126	113		
Eligible for free or reduced-price meals				
Mean	45.8	49.3	–3.5	.55
Standard deviation	34.9	35.7		
n	97	100		

a. *p*-value for quantitative data determined through Monte Carlo estimation of exact Wilcoxon rank sum test. *p*-value for categorical data determined through two-tailed Fisher's exact test.
Source: Author's analysis of primary data collected for the study.

Data analysis methods

Before impact analyses were conducted, comprehensive analysis files were created that included all outcome variables (confirmatory and exploratory); all subgroup variables; all covariates (with missing values replaced with the average of non-missing values, see below); student, teacher, and district identifiers (which did double duty as level/clustering variables in the hierarchical linear model analyses); and a single treatment variable, *Tx*, denoting the experimental treatment status. R version 2.9.2 (2009-08-24) software was used to conduct these analyses (R Foundation for Statistical Computing 2009).

Impact analyses

Multilevel regression models (also known as hierarchical linear models) were estimated to test the main research hypotheses. Adjusted post-intervention outcomes for students and teachers in the intervention group were compared with the outcomes for their counterparts in the control group. The primary hypothesis-testing analyses involved fitting conditional multilevel regression models with random intercepts to account for the nesting of individuals within higher units of aggregation (see, for example, Goldstein 1987; Murray 1998; Raudenbush and Bryk 2002). Statistically significant positive effects for hypothesis 1a, 1b,

2a, or 2b would constitute evidence supporting the effectiveness of the intervention. Statistically significant positive effects for hypothesis 3 or 4 would constitute evidence supporting the logic model for the intervention.

Multiple comparison procedures were used to adjust for inflation in the probability of false positive errors involving tests of the four hypotheses. Statistical results are reported with two sets of *p*-values, both adjusted for multiple comparisons and with no adjustments, to allow comparison with studies that do not include adjustments. Student-level results were adjusted for two comparisons (ATLAST and standardized achievement tests) for the full sample and the English language learner subsample. Teacher-level results were also adjusted for two comparisons (ATLAST tests and confidence levels). Results are considered statistically significant in this report only if the adjusted *p*-value is less than 0.05.

For purpose of the analysis, teachers were nested in randomization blocks within research sites; for student-level outcomes, students were nested within teachers. Students were not nested within classes within teachers, because a sensitivity analysis indicated that the additional level did not affect impact estimates (see appendix O). Covariates at the site, teacher, and student level were included in the analysis of student-level outcomes, and site-level and teacher-level covariates were included in the analysis of teacher-level outcomes (table 2.16). A detailed discussion of the model specification is in appendix P.

Table 2.16. Covariates included in student- and teacher-level regression models

Student-level model	*Teacher-level model*
Pretest	Pretest
Teacher randomization stratum	Teacher randomization stratum
Treatment group of teacher	Treatment group of teacher
Site-by-treatment interaction	Site-by-treatment interaction
Student sex	Class-level student academic ability
Student English language learner status	
Student race/ethnicity	
Teacher	
Teacher pretest	
Teacher sex	Teacher sex
Teacher bachelor's degree	Teacher bachelor's degree
Teaching experience	Teaching experience
	Teacher initial confidence
Missing-value indicators	Missing-value indicators
Error	Error

Source: Author.

Missing data

All data for the teacher and student analytic samples were examined to identify missing item-level responses (see appendix Q). Missing data rates were not statistically significantly different between the intervention and control group teachers, including for the ATLAST Test of Force and Motion.

Cases with missing values on covariates were retained in the analysis. In the context of a randomized controlled trial, where randomization helps ensure that the baseline covariates are balanced, the use of the missing indicator method appears to produce unbiased impact estimates and standard errors (Puma et al. 2009; White and Thompson 2005). Cases with missing values on posttest or other outcome variables were excluded from the impact analyses. Deletion of cases with missing outcome variables has been shown to result in accurate impact estimates and standard errors when outcomes are missing at random conditional on the covariates (Allison 2002; Puma, Olsen, Bell, et al. 2009; von Hippel 2007).

To deal with item-level missing covariate values, the research team created total scale scores by averaging non-missing values for that item (e.g., all missing pretest scores were coded to a constant, computed as the mean of non-missing pretest scores across all intervention and control group teachers). The mean was used so that those observations would have no weight on the estimate of the relationship between covariates and outcomes.

To account for missing values used in the impact analysis models, the research team used the missing-indicator method (White and Thompson 2005) wherein, in the HLM analyses, a missing-indicator categorical variable was set to zero or one (0 = observed; 1 = missing). Both the recoded covariates and the missing-value indicator variables were included in the regression model.

Sensitivity analyses were conducted to ascertain the stability of impact estimates using different samples and model specifications. Models with different combinations of covariates (no covariates, pretest measure of outcome variable only, or all covariates) were estimated on (a) the sample with valid data on the post-instruction outcome measures, (b) the sample with valid data on both pre-intervention and post-instruction outcome measures, and (c) the full sample. For samples (a) and (b), which included observations with non-missing outcome and/or covariate values, the missing-indicator method was used to estimate models across analytic samples.

Multiple comparison procedures

The procedures described by Schochet (2008) were used to account for multiple hypothesis tests involving the outcome variables assessed in the study. Two primary student outcomes were assessed: student content knowledge of force and motion and student content knowledge of physical science more generally. These outcomes were analyzed across two overlapping samples, the full sample of students in participating teachers' classrooms and the subsample of English language learners in participating teachers' classrooms. With two primary outcomes analyzed, adjustments for two statistical tests were applied to the impact estimates. The full sample of students was analyzed, with multiple comparison adjustments for two statistical tests. The subsample of ELL students was then analyzed separately, also with multiple comparison adjustments for two statistical tests.

Two intermediate teacher-level outcomes were analyzed to establish part of the theory of action linking the intervention to student academic skills and knowledge: teacher content knowledge of force and motion and teacher confidence in teaching ability. Multiple comparison procedures were used to adjust for the inflation of type I errors across the two statistical tests. Multiple comparisons were controlled for separately when analyzing the primary student outcomes and intermediate teacher-level outcomes.

Benjamini and Hochberg's (1995) stepwise multiple hypothesis testing procedure was used to adjust *p*-values. This procedure involves ordering *p*-values obtained for outcomes within each domain from largest to smallest, multiplying each unadjusted *p*-value by $N/(N - j + 1)$, where N is the number of outcome variables within a domain and j represents the order of the test. As applied in this study, all null hypotheses are rejected in which the adjusted *p*-value is less than 0.05.

Chapter 3. Implementation of the Making Sense of SCIENCE™ intervention

The professional development intervention consisted of a 24-hour force and motion course for teachers, delivered in summer 2009 over a period of five days (see appendix R). The course was implemented regionally, with local facilitators leading the course for local teachers at each of the six research sites.

Course content is divided into five sessions that are sequenced so that the science topics (for example, speed, velocity, acceleration, and balanced and unbalanced forces) build on one another. The corresponding science language issues and strategies for supporting student learning and language development are introduced incrementally over the sessions. Each session includes the four main components described in chapter 1 (hands-on science investigations, language and literacy activities, case discussions, and classroom connections).

Course materials

A teacher book was provided to teachers and a facilitator guide to staff developers. The teacher book contains five chapters (one per session) and presents all the materials teachers need to participate in the professional development course. Each chapter contains a teaching case that illustrates students' science thinking and highlights an important teaching dilemma that any teacher might face; a companion content guide explains and illustrates core science concepts. Each chapter in the teacher book also includes science investigation and case discussion handouts, which guide teachers' small-group working time and structure their conversations about science, student thinking, and instruction.

The facilitator guide contains five chapters (one per session). It provides extensive support materials and detailed procedures needed to successfully lead a course. Each chapter describes the underlying science (including common yet incorrect ideas children and adults have) and provides scripted yet flexible procedures, such as instructions to guide the hands-on and sense-making work in each science investigation, guiding questions for each case discussion, and instructions for helping teachers complete classroom connection assignments between sessions.

Facilitator selection and training

Site coordinators and district staff at each site helped identify and solicit the participation of professional development leaders to facilitate the courses. Understanding Science for Teaching staff participated in selecting the facilitators from among those individuals through telephone calls with candidates. In selecting facilitators, the research staff considered the following qualifications:

- At least two years' experience leading teacher professional development courses in middle school science.
- Strong science content knowledge, ideally college-level coursework in physical science including the specific content topic of the professional development course.

- At least five years of experience teaching the focal content to the grade addressed in the study.
- Strong pedagogical content knowledge, including ability to describe what tends to be difficult for students and teachers to understand about force and motion and ability to generate instructional strategies that address specific learner misconceptions.
- Good fit with the Making Sense of SCIENCE^{TM} professional development model, including a social constructivist perspective focusing on helping students and teachers learn through collaborative discourse about science.
- Acceptance of and commitment to following a strict professional development and research protocol for the larger good in science education.

Two facilitators were selected from each site. In July 2009 all facilitators were trained to lead the course in one five-day leadership academy held at WestEd in Oakland, California. Facilitators were introduced to the purpose and design of the research and experienced the professional development intervention themselves. Most of the training time was spent deepening facilitators' understanding of force and motion, grounding them in the common yet incorrect ideas students (and adults) have, and helping participants develop the necessary facilitation skills. Project staff modeled facilitation, engaged the group in analyzing video clips of exemplary facilitation, and provided facilitation course sessions for local teachers who volunteered to participate in the practice sessions. Facilitators used the course materials—the facilitator guide and participant book—throughout the training.

Course implementation

In summer 2009 teachers who were randomly assigned to the intervention group took the Understanding Force and Motion course, led by pairs of trained facilitators at each site, who alternated between serving as lead facilitator and serving as co-facilitator for each session. An average of more than 80 percent of teachers initially assigned to the intervention group received the intervention, ranging from 73 percent to 100 percent at individual sites (table 3.1). 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. Among intervention group teachers still in the study, 94 percent attended the course, ranging from 85 percent to 100 percent at individual sites.

It should be noted that teachers in both the intervention and control groups were permitted to participate in other professional development besides the courses under investigation in the present study. Data as to such participation are available from the teacher survey, but worries about the quality of these data precluded us from presenting them. Response patterns suggest that teacher respondents in the treatment group did not distinguish between the professional development they received via Making Sense of SCIENCE from that received from other sources.

Table 3.1. Number of teachers assigned to and participating in summer 2009 Making Sense of SCIENCE™ courses, by research site

Site	*Percentage of initially assigned teachers completing course*	*Percentage of available teachers completing course*[a]
1	78.6	84.6
2	84.6	100.0
3	73.3	91.7
4	73.3	84.6
5	77.8	100.0
6	100.0	100.0
Total	81.1	93.6

a. Excludes teachers who could not participate because study was not approved at district or school level and teachers who left teaching or moved to a different grade or school.
Source: Author's analysis of primary data collected for the study.

Two video recordings, chosen at random, were reviewed at each site to monitor fidelity of implementation (see appendix R). Review of the 12 sessions revealed perfect correspondence between the course components as designed and as implemented. Debriefing conversations between course facilitators and Understanding Science for Teaching program staff indicated that no facilitators dropped a course component included in the facilitator guide. Records kept for each session of each course indicated nearly perfect attendance.

Cost of training teachers in Making Sense of SCIENCE™

The estimated cost of providing the five-session courses to intervention group teachers at the six research sites was $107,900 (table 3.2). This figure includes materials, training, logistical supports, and reimbursement of teachers for professional time. It reflects the fact that the professional development workshops were held at school sites or other locations that did not require a facility rental fee. Had these sites not been available, an additional $4,500 ($150/session × five sessions × six sites) would have been needed.

Table 3.2. Estimated cost of training teachers in Making Sense of SCIENCE™

Item	Estimated unit cost (dollars)	Number of units	Estimated total cost
Teacher stipend	800/teacher	73 (12 teachers/site × 6 sites)	$58,400
Facilitator stipend	1,800/ facilitator	12 (2 facilitators/site × 6 sites)	$21,600
Facilitator training (travel expenses for five-day training)	1,500/facilitator	12	$18,000
Hands-on materials	200/site	6	$1,200
Curricular materials for teachers	60/teacher	73	$4,380
Curricular materials for facilitators	360/facilitators	12	$4,320
Total			$107,900

Source: Author's analysis of primary data collected for the study.

Implementation at the classroom level

The intervention evaluated is not a student curriculum but a teacher course designed to improve students' learning opportunities by strengthening teachers' science and pedagogical knowledge. Making Sense of SCIENCE™ courses are intended to strengthen teaching in a way that is compatible with whatever student curriculum is already used in the classroom. No materials were provided for use in teachers' classrooms, although some teachers did adapt activities they completed in the course for student use.

Curriculum decisions are made at the district or school level at least a year in advance of each school year. If teachers' experience in the course changes the instructional methods or approaches they want to use, it is more likely that they will supplement their regular curriculum from other resources than that they will change curricula overall.

Although change in student curricula in the classroom as a result of the course is neither intended nor likely, determining whether it may have occurred is important. If teachers who took the course changed the textbook or curriculum they used as a result, then course impact could be largely a function of the book used rather than how the books and other materials were used. Questions on teacher surveys solicited information about student curricula used the year before the Making Sense of SCIENCE™ course was given and during the study year. In both years teachers in the intervention and control groups used the same set of curricula, as would be expected with random assignment within schools and districts (table 3.3).

Table 3.3. Science textbooks used by teachers before and during study year, by experimental condition and curriculum

	2008/09				*2009/10*			
	Intervention group		*Control group*		*Intervention group*		*Control group*	
Publisher	*Number*	*Percent*	*Number*	*Percent*	*Number*	*Percent*	*Number*	*Percent*
Pearson Prentice Hall	21	26.9	21	28.4	15	20.3	14	23.0
Holt, Rinehart and Winston	12	15.4	18	24.3	13	17.6	16	26.2
CPO Science	12	15.4	10	13.5	9	12.2	7	11.5
Herff Jones	12	15.4	6	8.1	12	16.2	6	9.8
Glencoe/McGraw Hill	7	9.0	#	#	8	10.8	5	8.2
District materials	6	7.7	#	#	6	8.1	#	4.9
Other	8	10.2	11	15.0	11	15.0	#	16.4
Total	78	100	74	100	74	100	61	100

Notes: Numbers of teachers may not sum to expected sample sizes because some teachers reported using more than one student curriculum. Components may not sum to 100 because of rounding. *P*-values, calculated using two-tailed Fisher's exact test for equality of proportion between intervention and control group teachers, were 0.52 for 2008/09 and 0.81 for 2009/10.

# indicates values were suppressed to reduce disclosure risk.

Source: Author's analysis of primary data collected for the study.

The distributions of science textbooks between the two groups of teachers did not differ statistically either year. Thus, there is no evidence that the Making Sense of SCIENCE^{TM} course prompted changes in textbook by intervention group teachers.

Chapter 4. Impact results

Results of the primary confirmatory analyses indicate that, after adjusting for multiple comparisons, there were no statistically significant differences between science content test gains of students in intervention group classrooms and students in control group classrooms. Intervention group students in neither the full sample nor the English language learner subsample scored significantly higher than control group students on the ATLAST Test of Force and Motion or the California Standards Test physical science reporting clusters. These findings reflect a crucial element of the analysis plan, namely that two null hypotheses were to be tested, and therefore adjustments were required for two comparisons (intervention versus control outcomes on both the ATLAST test and the California Standards Test physical science reporting clusters scores).

Results of the intermediate confirmatory analyses indicate that after adjusting for multiple comparisons, teachers who took the Making Sense of SCIENCE^{TM} course outscored control group teachers on the ATLAST Test of Force and Motion for Teachers (effect size = 0.38). They also revealed higher self-ratings of confidence in the ability to teach force and motion (effect size = 0.49).

Student outcomes (primary research questions)

Evidence on hypotheses 1a and 1b: Did Making Sense of SCIENCE^{TM} professional development increase students' content knowledge of force and motion or of physical science more generally?

After adjustment for multiple comparisons, differences between the full sample of students in the intervention and control groups on the ATLAST Test of Force and Motion were not statistically significant at the 0.05 level (table 4.1). Intervention group students in the full sample (effect size = 0.11) did not score higher than control group students on the ATLAST Test of Force and Motion.

Similarly for the California Standards Test physical science reporting clusters, intervention group students in the full sample (effect size = 0.03) did not score higher than their counterparts in the control group.

Table 4.1. Impact analysis of science content knowledge outcomes for all students

Measure	Adjusted mean: Intervention group (standard deviation)	Adjusted mean: Control group (standard deviation)	Difference (standard error)	Unadjusted p-value	Statistical significance after correction[a]	Effect size	Student sample size
ATLAST Test of Force and Motion (percent correct)	52.4 (19.8)	50.3 (19.3)	2.1 (1.0)	.04	No	0.11	5,130
California Standards Test physical science reporting clusters	71.0 (19.3)	70.4 (18.2)	0.5 (1.1)	.65	No	0.03	3,768

ATLAST is Assessing Teacher Learning About Science Teaching.
Notes: Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.
a. Benjamini-Hochberg correction used to adjust for multiple comparisons of two outcomes.
Source: Author's analysis of primary data collected for the study.

Evidence on hypotheses 2a and 2b: Did Making Sense of SCIENCE™ professional development increase English language learners' content knowledge of force and motion or of physical science more generally?

After adjustments for multiple comparisons, differences between English language leaner students in the intervention and control groups on the ATLAST Test of Force and Motion were not statistically significant at the 0.05 level (table 4.2). The sample of intervention group English language learners did not outscore the sample of control group English language learners on the California Standards Test physical science reporting clusters.

Table 4.2. Impact analysis of science content knowledge outcomes for English language learner students

Measure	Adjusted mean: Intervention group (standard deviation)	Adjusted mean: Control group (standard deviation)	Difference (standard error)	Unadjusted p-value	Statistical significance after correction[a]	Effect size	Student sample size
ATLAST Test of Force and Motion (percent correct)	38.9 (15.3)	34.4 (14.2)	4.5 (2.2)	.04	No	0.31	455
California Standards Test physical science reporting clusters	54.7 (19.2)	56.5 (19.7)	–1.8 (2.6)	.50	No	0.09	378

Note: Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.
ATLAST is Assessing Teacher Learning About Science Teaching.
a. Benjamini-Hochberg correction used to adjust for multiple comparisons of two outcomes.
Source: Author's analysis of primary data collected for the study.

Teacher outcomes (intermediate research questions)

Evidence on hypothesis 3: Did Making Sense of SCIENCE™ professional development increase teachers' content knowledge of force and motion?

The intervention increased teachers' content knowledge of force and motion, as measured by the ATLAST Test of Force and Motion (table 4.3). Adjusted mean differences on the posttest measure in spring 2010 were 6.2 percentage points higher for the intervention group (effect size = 0.38). This difference was significant at the 0.01 level after adjusting for multiple comparisons across two teacher-level domains using the Benjamini-Hochberg (1995) procedure.

Table 4.3. Impact analysis of teacher science content knowledge and confidence in ability to teach force and motion

	Adjusted mean						
Measure	*Intervention group (standard deviation)*	*Control group (standard deviation)*	*Difference (standard error)*	*Unadjusted p-value*	*Statistical significance after correction*[a]	*Effect size*	*Teacher sample size*
ATLAST Test of Force and Motion (percent correct)	65.3 (19.2)	59.2 (16.0)	6.2** (2.2)	<.01	Yes	0.38	133
Confidence in ability to teach force and motion[a]	2.7 (0.3)	2.5 (0.4)	0.2** (0.04)	<.01	Yes	0.49	133

Notes: Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.
ATLAST is Assessing Teacher Learning About Science Teaching.
*Significantly different from zero at the 0.05 level, two-tailed test. **Significantly different from zero at the 0.01 level, two-tailed test.
a. Benjamini-Hochberg correction used to adjust for multiple comparisons of two outcomes.
b. Based on teacher ratings on a three-point Likert scale that ranged from 0 (not at all confident) to 3 (very confident)
Source: Author's analysis of primary data collected for the study.

Evidence on hypothesis 4: Did Making Sense of SCIENCE™ professional development increase teachers' confidence in their ability to teach force and motion?

The intervention produced gains in teachers' confidence in their ability to teach force and motion (see table 4.3). Adjusted mean differences on the confidence measure in the spring 2010 semester show that the outcome for the intervention group exceeded that of the control group (confidence rating effect estimate of 0.2; effect size = 0.49). This difference was significant at the 0.01 level after adjusting for multiple comparisons across two teacher-level domains.

Sensitivity analyses

All primary analyses were conducted using impact models estimated with a full set of relevant covariates; for samples with valid, non-missing posttests; and with any missing pretest and covariate values replaced with the average of non-missing values. The robustness of treatment effects was examined by determining the sensitivity of findings to models estimated with different combinations of covariates and different analytic samples (see appendix S).

Influence of student-level covariates and analytic student sample

Estimates of impacts on the ATLAST Test of Force and Motion or the California Standards Test physical science reporting clusters were similar whether or not covariates were included in the models. There was also very little variation when different analytic samples were used.

Influence of teacher-level covariates

Treatment effects were estimated for the same three sets of covariates that were compared for students (all with $n = 133$). All models were estimated for the teacher sample with valid, non-missing posttests. Treatment effects on teachers' content knowledge of force and motion reached statistical significance for all three models. However, the inclusion of the pretest in the impact analysis model decreased the point estimate from 9.8 to 6.1 and the effect size from 0.61 to 0.38. The differences in estimates when the pretest was included in the basic model likely reflected the significant differences between baseline science scores of intervention and control group teachers (see table 2.5). There were no differences between estimates for the model with pretest only and estimates for the model with all covariates.

With regard to treatment effects on teachers' confidence in their ability to teach force and motion, controlling for covariates did not significantly change the outcome. Treatment effects on confidence reached statistical significance for all three models, with effect sizes of 0.46–0.49.

Influence of analytic teacher sample

Estimating effects for different analytic samples did not change the outcome with regard to teachers' content knowledge or confidence in their ability to teach force and motion. Treatment effects on teachers' content knowledge of force and motion reached statistical significance for all three models. There were no differences between point estimates (6.2), p-values (0.05), or effect sizes (0.38) for the models with additional missing values. Treatment effects on teacher confidence reached statistical significance for all three models ($p < .01$, effect size = 0.49).

Chapter 5. Exploratory analyses

According to the Understanding Science for Teaching program's theory of action, increased teacher content knowledge is a key intermediate outcome of the teacher courses. If course implementation at each site differentially affects teacher content knowledge, student-level effects would be expected to be similarly affected. Therefore, the study team explored the relationship between student and teacher outcomes by examining whether the pattern of student and teacher impacts varied across the six implementation sites.

These analyses focused only on content knowledge of force and motion, as measured by the ATLAST tests, in order to minimize multiple testing issues and because intervention-control differences were at or near significance for both teachers and students on these measures. The exploratory questions were addressed for the teacher sample and the full student sample only, because the site-level sample sizes for English language learners were too small to yield reliable results.

Differential impacts across sites

Did the impacts of Making Sense of SCIENCE^{TM} on student and teacher outcomes differ significantly across the six implementation sites? What, if any, were the differential impacts by site of the course on students' content knowledge of force and motion?

To address these questions, the study team analyzed teacher and student data including treatment-site interaction variables. These analyses indicated whether there were differential treatment effects across sites and allowed separate impacts to be estimated for the sites. A likelihood ratio test comparing the results of these models with the main impact analysis models that did not include the treatment-site interaction terms was used to determine whether there were statistically significant differences between models with and without site-by-treatment interactions.

Results of the likelihood ratio tests indicated that neither student nor teacher models differed, suggesting that the intervention effects did not vary by site. The model that included site-by-treatment interaction terms, site-specific impact estimates, *p*-values, and effect sizes showed the most pronounced effects for two of the six sites, for both students (table 5.1) and teachers (table 5.2). An important limitation of the exploratory analyses is the limited statistical power for estimation of site-specific impacts (see appendix A).

Table 5.1. Impact analysis of student content knowledge of force and motion, by site

	Adjusted mean					*Unweighted*	
Site	*Intervention group*	*Control group*	*Difference*	p-*value*	*Confidence interval*	*Effect size*	*Student (teacher) sample size*
1	50.9	45.8	5.0*	.03	0.4 to 9.6	0.26	848 (23)
2	57.5	55.6	2.0	.47	–3.5 to 7.4	0.10	706 (14)
3	52.0	48.2	3.8	.08	–0.6 to 8.1	0.19	1,027 (24)
4	56.2	55.9	0.3	.91	–5.4 to 6.1	0.02	641 (16)
5	51.8	50.0	1.8	.45	–2.9 to 6.5	0.09	1,032 (25)
6	48.7	49.2	–0.5	.83	–5.2 to 4.2	–0.03	867 (25)

Note: Results are based on student scores on the Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion. Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.
*Significantly different from zero at the .05 level, two-tailed test.
Source: Author's analysis of primary data collected for the study.

Table 5.2. Impact analysis of teacher content knowledge of force and motion, by site

	Adjusted mean					*Unweighted*	
Site	*Intervention group*	*Control group*	*Difference*	p-*value*	*Confidence interval*	*Effect size*	*Teacher sample size*
1	63.2	50.8	12.4*	.02	1.6 to 23.2	0.77	23
2	63.8	55.7	8.1	.17	–3.8 to 20.0	0.51	17
3	68.5	57.1	11.4*	.03	1.1 to 21.8	0.71	24
4	75.7	72.9	2.9	.63	–9.0 to 14.7	0.18	19
5	67.1	60.8	6.4	.25	–4.7 to 17.4	0.40	25
6	55.2	59.3	–4.1	.46	–15.2 to 7.0	–0.26	25

Note: Results are based on student scores on the Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion. Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.
*Significantly different from zero at the .05 level, two-tailed test.
Source: Author's analysis of primary data collected for the study.

How do the patterns of and differences in impacts across sites for teacher outcomes compare with those for student outcomes?

Although statistical analysis of correlations between teacher and student ATLAST test score outcomes is not advisable with only six implementation sites, examination of the estimated treatment effects for teachers and students at each site reveals a distinct pattern (table 5.3). Point estimates of student content knowledge of force and motion and teacher content knowledge of force and motion follow the same rank order, without exception. This pattern is consistent with a relationship between teacher and student outcomes, but without significant student-level effects this relationship could not be investigated.

Table 5.3. Impact point estimates for knowledge of force and motion by teachers and students

Site	*Teacher knowledge* (*n* = 133)	*Student knowledge* (*n* = 5,130)
1	12.4	5.0
3	11.4	3.8
2	8.1	2.0
5	6.4	1.8
4	2.9	0.3
6	–4.1	–0.5

Note: Knowledge of force and motion was measured by the Assessing Teacher Learning About Science Teaching (ATLAST) Test of Force and Motion for Students and the ATLAST Test of Force and Motion for Teachers.
Source: Author's analysis of primary data collected for the study.

Chapter 6. Conclusion

Primary confirmatory analyses at the student level indicate that after adjusting for multiple comparisons, there were no statistically significant differences between science content test scores of students whose teachers participated in the Making Sense of SCIENCE™ professional development course on force and motion and students in control group classrooms, at least at conventional levels of statistical significance.

Results for the intermediate confirmatory analyses at the teacher level indicate that after adjusting for multiple comparisons, teachers who received the course outscored control group teachers on the ATLAST Test of Force and Motion for Teachers ($p < 0.01$, effect size = 0.38) as well as in their ratings of confidence in their ability to teach force and motion ($p < 0.01$, effect size = 0.49).

The estimated impacts were consistent when tested using models estimated with different combinations of covariates and different analytic samples. Estimating effects for different covariates and samples did not significantly change the outcomes with respect to treatment effects on students' or teachers' content knowledge of force and motion or teacher confidence.

In exploratory analyses, the study team examined whether there were differential impacts on student and teacher content knowledge outcomes across the six research sites. It found that the estimated impacts on student and teacher content knowledge of force and motion were greatest at two of the six sites. The relationship between student effects and teacher effects followed the same pattern at the six sites, with the rank order of student effects exactly matching the rank order of teacher effects. The finding could mean that at sites at which the intervention was particularly effective, teachers learned many things (including content knowledge) and that students gained more because of a combination of treatment effects. In this case, the average increase in the content knowledge scores related to treatment might be some measure of the sites' overall implementation of the intervention. An important limitation of this exploratory analysis is the weak statistical power for estimation of site-specific impacts.

To examine further whether treatment effects on student outcomes were mediated by teacher content knowledge gains, the study team controlled for teacher posttest scores in student-level hierarchical linear models. Doing so reduced the student-level impact estimate by just 6 percent. Although teacher content knowledge may mediate student impact, these findings suggest that, as represented in the intervention logic model, the course produces student gains by influencing more than just teacher content knowledge outcomes.

Implications of the results

The analysis plan for this study established that statistically significant positive effects for any of the two hypotheses involving student knowledge would constitute evidence supporting the effectiveness of the Making Sense of SCIENCE™ teacher professional development intervention and that only evidence related to the two hypotheses would be used to make inferences about the overall effectiveness of the program. For both the full student sample and the sample of English language learners, neither of the two null hypotheses was rejected at

$p < 0.05$ after adjustments for the two comparisons. The findings therefore are inconclusive with respect to the effectiveness of the intervention.

At the teacher level, treatment effects were clearly positive and significant: the findings support the effectiveness of the intervention for raising teacher science content test scores. These results are consistent with all previous evaluations of Making Sense of SCIENCE™ teacher courses (Heller, Daehler, and Shinohara 2003; Heller et al. 2010), signaling to educators and policymakers that the Making Sense of SCIENCE™ force and motion course can be relied on to strengthen the science content knowledge of teachers.

In exploratory analyses, the study investigated whether there were differential impacts on student and teacher content knowledge outcomes across the six research sites. The estimated impacts were most pronounced at two of the six sites. For the full sample of students, point estimates for student and teacher content knowledge of force and motion followed the same rank order at all sites.

Limitations of the analysis

As described in chapter 2, 48 of the 181 teachers who were randomly assigned to intervention and control groups left the study before data collection was completed, raising concerns about attrition bias. To the extent that these teachers differed from participating teachers, such attrition could reduce external validity (the degree to which the results can be generalized from the retained teacher sample). Such attrition could also bias impact estimates if the attrition is associated with the study outcome measures and attrition rates differ between intervention and control groups (What Works Clearinghouse 2008). Based on the analyses of equivalence between the intervention and control groups at baseline and at subsequent points later in the study, as well as between retained and nonretained teacher samples, there is little evidence of selective attrition. Sensitivity analyses conducted (reported in appendix S) also show consistent findings with analytic samples based on missing data as a result of participant attrition and unresponsiveness to data collection protocols.

Another limitation of the study is that data were not collected on classroom implementation of course-related practices, which might help to explain the absence of student-level effects. The expense of collecting extensive classroom implementation data weighed against conducting a detailed process study, without which it is not possible to determine whether the course affected teachers' practices.

The findings are based on volunteer teachers and students whose parents provided consent. It is possible that the findings would have been different had teachers been required to participate in the intervention and all students been tested.

Appendix A. Study power estimates

This appendix describes how the sample sizes were chosen for this study.

Power estimates during planning phase

To determine the appropriate sample sizes, during the planning phase the study team calculated minimum detectible effect sizes based on the unit of randomization, the sources of clustering, the availability of baseline explanatory variables, and other design characteristics, using the procedures described by Donner and Klar (2000), Murray (1998), Raudenbush (1997), and Schochet (2005). Minimum detectible effect size estimates represent the smallest true program impacts (in standard deviation units) that can be detected with high probability (Bloom 1995). The minimum detectible effect size of a study is the smallest effect size that has at least an 80 percent probability of being found statistically significant with 95 percent confidence. For a design to be sufficiently powerful, this minimum detectible effect size must be small enough so that a likely program impact that is large enough to be policy relevant does not go undetected.

Fourteen parameters were used to estimate minimum detectible effect size (table A1). As discussed in the body of this report, the study team estimated that 120 of approximately 180 teachers randomly assigned to two conditions would be retained after attrition; that each teacher would cover two classes with about 25 students per class; that the student nonresponse/missing-data rate would be about 20 percent, leaving 20 students per class and 40 students per teacher at the end of the semester for analysis; and that 25 percent of student participants served by each teacher would be classified as English language learners.

Table A1. Parameters used to estimate statistical power in planning phase and actual parameters in final analytic sample

	Planning phase		*Final analytic sample*	
Parameter	*Student outcomes*	*Teacher outcomes*	*Student outcomes*	*Teacher outcomes*
Teachers				
Teachers per condition	90	90	90	90
Participating teachers per condition	60	60	65[a]	66[a]
Participating teachers per condition in California[b]	50	na	53	na
Students				
Students per teacher	50	na	61	na
Participating students per teacher	40	na	40	na
Participating English language learners per teacher	10	na	3	na
Intraclass correlation				
ATLAST Test of Force and Motion for Students	0.20	na	0.19	na
California Standards Test physical science reporting clusters	0.20	na	0.35	na
R^2 *(within-teacher)*				
ATLAST Test of Force and Motion for Students	0.50	na	0.34	na
Student California Standards Test physical science reporting clusters	0.50	na	0.29	na
R^2 *(between-teacher)*				
ATLAST Test of Force and Motion for Students	0.50	na	0.73	na
California Standards Test physical science reporting clusters	0.50	na	0.86	na
ATLAST Test of Force and Motion for Teachers	na	0.20	na	0.74
Teacher confidence	na	0.20	na	0.74

Note: All parameters except the number of teachers per condition and the number of students per teacher were used to estimate minimum detectable effect size.
na is not applicable.
ATLAST is Assessing Teacher Learning About Science Teaching.
Harmonic mean of the number of teachers in each experimental condition.
Student state standardized test score information collected only for California sample.
Source: Author's analysis of primary data collected for the study.

For the purposes of the power analyses, the study team conservatively assumed intraclass correlations of 0.20 for the student academic outcomes and between- and within-teacher R^2 values of 0.50, based on Schochet's (2005) work. Based on other studies of teacher outcomes (for example, Hill and Ball 2004; Schweingruber and Nease 2000), it conservatively assumed that covariates would explain 20 percent of the variance in teacher outcomes. Using a Bonferroni adjustment as a conservative approximation of the proposed

resampling method, the study team divided the critical value of the statistical significance test by four for the primary student outcomes and by two for the intermediate teacher outcomes.

With 60 teachers per condition and a minimum of 40 (25 × 2 × 0.80) students and 10 English language learner students per teacher, the study team estimated the minimum detectible effect size to be 0.20 for ATLAST test scores involving the total student sample and 0.23 for the English language learner sample (table A2). As noted in the body of this report, standardized test score information was collected only from the California sample, which included about 50 teachers per condition. The estimated minimum detectible effect size for standardized test scores on the physical science reporting clusters of the California Standards Test was 0.22 for the full sample, 0.25 for the English language learner subsample, and 0.51 for the teacher outcomes.

Table A2. Minimum detectable effect size estimates for student and teacher outcome measures

Sample	*Planned sample minimum detectible effect size*	*Achieved sample minimum detectible effect size*
All students		
ATLAST Test of Force and Motion for Students	0.20	0.15
California Standards Test physical science reporting clusters	0.22	0.15
English language learner students		
ATLAST Test of Force and Motion for Students	0.23	0.28
California Standards Test physical science reporting clusters	0.25	0.27
Teachers		
ATLAST Test of Force and Motion for Teachers	0.51	0.28
Teacher confidence	0.51	0.28

Note: Calculations assumed type I error rates of 0.05 (two-sided) and a fixed-effects statistical model. See table A1 for other parameters used to estimate minimum detectable effect sizes.
ATLAST is Assessing Teacher Learning About Science Teaching.
Source: Author's analysis of primary data collected for the study.

Power estimates for final analytic sample

Greater numbers of teachers participated in the study than anticipated during planning. The final analytic sample included 133 teachers providing teacher survey data and 131 teachers providing student data. On average, data were eligible for analysis for 40 students per teacher (5,251 students total with at least posttest data on the ATLAST Test of Force and Motion). The intraclass correlations for the student outcomes were 0.19 for the ATLAST scores and 0.35 for the scores on the California Standards Test physical science reporting clusters (see table A1).

The estimated within-teacher R^2 values were smaller than anticipated at the planning stage, and the between-teacher R^2 values were larger than expected. The greater than expected number of teachers participating, combined with higher R^2 values, resulted in statistical power gains for the overall student sample, with a minimum detectible effect size of 0.15 for the ATLAST and standardized test scores (see table A2). The minimum detectible effect size for teacher intermediate outcomes was 0.28—substantially lower than estimated during the study planning stage because of the larger proportion of variation explained by covariates than originally assumed.

Fewer English language learner students were available for analysis than anticipated, with an average of 3 (rather than 10) students per classroom with valid data. For the English language learner subsample, the minimum detectible effect sizes were 0.28 for the ATLAST Test of Force and Motion and 0.27 for the standardized test scores.

Power estimates for exploratory analyses

The parameters listed in table A1 for the achieved sample were used to estimate minimum detectible effect size estimates for site-specific impacts. These estimates were based on 11 teachers per condition at each site and made no adjustments for multiple hypothesis tests (table A3). The site-specific minimum detectible effect size estimates for student academic outcomes were 0.31–0.32 for the overall student sample, 0.56–0.60 for the English language learner student subsample, and 0.64 for teacher outcomes 0.64.

Table A3. Site-specific minimum detectable effect size estimates for student and teacher outcome measures

Group/measure	*Minimum detectible effect size (standard deviations)*
All students	
ATLAST Test of Force and Motion for Students	0.32
California Standards Test physical science reporting clusters	0.31
English language learner students	
ATLAST Test of Force and Motion for Students	0.60
California Standards Test physical science reporting clusters	0.56
Teachers	
ATLAST Test of Force and Motion for Teachers	0.64
Teacher confidence	0.64

ATLAST is Assessing Teacher Learning About Science Teaching.
Note: Calculations assumed 11 teachers per condition at each site, type I error rates of 0.05 (two-sided), and a fixed-effects statistical model. See table A1 for the other parameters used to estimate minimum detectable effect sizes.
Source: Author's analysis of primary data collected for the study.

Appendix B. Procedure for assigning blocks for recruited sample and final analytic sample

The recruitment process required a random assignment design both within and between schools because, within each of the six research sites, there were two groups of teachers: one group from schools with two or more participating teachers and another group from schools with only one participating teacher. For schools with two or more participating teachers, the study team conducted the randomization within each school. Schools with only one participating teacher were first ranked based on 2008 school-level state test scores.[8] The ranked list was then separated into blocks consisting of two teachers each. The first teacher in each block was randomly assigned to either the intervention or the control group, the second to the other group. This procedure was followed at each regional site (table B1). It resulted in two kinds of randomization blocks at each site:

- Teacher-level blocks, each consisting of two teachers who were the only participants at their schools (or three teachers if there was an odd number of teachers at a site). At least one of these teachers was assigned to the intervention group and at least one to the control group. The assignment procedure generated 50 teacher-level blocks (48 blocks with two teachers and 2 blocks with three teachers).
- School-level blocks, each consisting of a school that had more than one teacher participant. Schools had at most three participating teachers, so these blocks included two or three teachers. At least one of these teachers was assigned to the intervention group and at least one to the control group. The assignment procedure generated 26 blocks for schools with two participating teachers and 9 blocks for schools with three participating teachers.

[8] In California the 2008/09 school-level mean percentages of students scoring at or above proficient on the grade 8 California Standards Tests of mathematics and reading was used to stratify schools. For schools at the Arizona site, the 2008/09 school-level mean scale scores on the grade 8 Arizona Instrument to Measure Standards in mathematics and reading were used.

Table B1. Numbers of teacher-level and school-level randomization blocks, by site

	Teacher-level blocks		*School-level blocks*		
Site	*Number of blocks with two teachers*	*Number of blocks with three teachers*	*Number of blocks with two teachers*	*Number of blocks with three teachers*	*Total blocks*
1	5	0	8	1	14
2	9	0	3	1	13
3	7	0	5	2	14
4	7	1	2	3	13
5	13	0	2	2	17
6	7	1	6	0	14
Total	48	2	26	9	85

Source: Author's analysis of primary data collected for the study.

By the end of the study, some blocks had changed, because of attrition, creating two additional situations:

- Singletons, consisting of only one teacher because the other teachers were no longer in the study.
- Blocks that still had two teachers remaining but in which both teachers were now in the same condition.

These situations are problematic when the variables for "experimental condition" and "block" are both included in the impact analysis models. Additional blocks were created to solve this problem. Within each site, "orphans" (teachers who had lost their partners) were pooled into a new block. At two of the sites, this generated a new block with all intervention group teachers or all control group teachers. At those sites, all the blocks were merged to form a sitewide stratum (which is the same as a site dummy variable for that site).

Appendix C. Teacher agreement to protect the study

Teacher Agreement to Protect the Study

Force and Motion for Teaching *2009/10*

Dear Colleague,

Thank you for volunteering to be part of this scientific test of a science professional development. Your cooperation in making this a valid study is extremely important and greatly appreciated. Please join us in making the study useful by making a commitment to protect it from threats to validity.

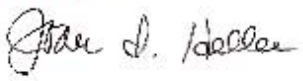
As you know, the purpose of this research is to compare outcomes of a WestEd professional development course with outcomes for teachers who have not yet taken the course. One of the serious challenges of a randomized study like this is that the results are **only** useful if the experiences of teachers participating in different groups are distinctly different. In this study, teachers are participating either by taking the *Force and Motion for Teaching* course or by completing the professional development in which they would ordinarily participate.

This agreement describes ways that you can help protect the study from "contamination" across groups as well as other threats to the study's validity. We are requesting that all teachers read and sign this agreement so that we may demonstrate to audiences of this research that we made every effort to conduct a sound, rigorous experiment.

Protecting our study from contamination is particularly crucial in schools and districts where teachers work closely with other teachers in the study who are not in the same group. **We must ask you not to spontaneously share or ask for detailed information about course activities, or course-related science content knowledge and pedagogical strategies, with your colleagues until the data have been collected in Spring 2010.** Thank you in advance for your commitment.

If you have any questions, please feel free to contact me directly.

Sincerely,



Joan I. Heller, Ph.D.
510-873-0800
jheller@edservices.org

KEEP THIS COPY

Teacher Agreement to Protect the Study

Force and Motion for Teaching *2009/10*

I agree to protect the differences between the two study conditions.

- ☐ I understand that giving or receiving *Force and Motion for Teaching* course materials to or from other teachers before Spring 2010 will compromise the research study and could jeopardize the effort that I and other teachers have given to this project.
- ☐ I will not ask other teachers for details about the *Force and Motion for Teaching* course until I am taking the course myself.
- ☐ I will only talk about or share details of *Force and Motion for Teaching* materials, activities, or approaches with teachers who have taken or are currently taking the *Force and Motion for Teaching* course.

I agree to protect the validity of students' performance on quizzes about force and motion.

- ☐ I will ask a colleague to proctor the student quizzes and I will not assist students during the quiz administration, except as noted in the administration instructions.
- ☐ I agree not to view the quizzes prior to, during, or after administering them to the students. I will return to Heller Research Associates all copies of the student quiz, and I will not copy or reproduce any part of the quizzes.

By signing the last page, I indicate that I am aware of, and agree to, these specific ways that I can support the study's validity.

SIGN AND RETURN THIS COPY

Teacher Agreement to Protect the Study

Force and Motion for Teaching *2009/10*

I agree to protect the differences between the two study conditions.

- ☐ I understand that giving or receiving *Force and Motion for Teaching* course materials to or from other teachers before Spring 2010 will compromise the research study and could jeopardize the effort that I and other teachers have given to this project.
- ☐ I will not ask other teachers for details about the *Force and Motion for Teaching* course until I am taking the course myself.
- ☐ I will only talk about or share details of *Force and Motion for Teaching* materials, activities, or approaches with teachers who have taken or are currently taking the *Force and Motion for Teaching* course.

I agree to protect the validity of students' performance on quizzes about force and motion.

- ☐ I will ask a colleague to proctor the student quizzes and I will not assist students during the quiz administration, except as noted in the administration instructions.
- ☐ I agree not to view the quizzes prior to, during, or after administering them to the students. I will return to Heller Research Associates all copies of the student quiz, and I will not copy or reproduce any part of the quizzes.

By signing here, I indicate that I am aware of, and agree to, these specific ways that I can support the study's validity.

_____
Printed Name

_____ _____
Signature Date

Appendix D. Teacher survey responses related to contamination across groups

Table D1. Teacher responses to end-of-year survey questions related to contamination across groups, for sample that was retained, by experimental condition

Measure	*Intervention*	*Control*
31a. To the best of your knowledge, have any teachers who did not participate in the WestEd Force and Motion for Teaching course begun to implement any aspects of that course?		
1. Yes	5.8%	6.5%
2. No	94.2%	93.6%
N	69	62
31b. If yes, how many teachers?		
Mean	1.3	1.3
SD	.5	.5
Range	1–2	1–2
N	69	64

Source: Author.

Appendix E. Parent consent form

Parent Consent Form

Force and Motion for Teaching *2009/10*

Fall 2009

Dear Parent or Guardian:

Your child's school is participating in an important research project to improve student science achievement. Your child's science teacher, [*NAME*], has chosen to be a part of this study and the principal of your school and the superintendent of your school district have formally reviewed and approved the study.

This letter is to introduce you to the study, which is called *Force and Motion for Teaching*. We explain how it involves your child, and ask for your consent that your child may participate.

The study is funded by the U.S. Department of Education and offers teachers advanced training in the field of science. We hope you will permit your child to take part in this exciting project along with all of his/her classmates. Naturally, all information gathered from students is kept strictly confidential.

Please fill out and sign the form on the last page, and return it to your child's science teacher within two weeks. I would be happy to answer any questions you may have. Your child's science teacher is also a good source of information. Thank you for your time and consideration.

Sincerely,

Joan I. Heller, Ph.D.
510-873-0800 ext. 1
jheller@edservices.org

Study Description

Force and Motion for Teaching *2009/10*

How the study will involve your child

Your child's teacher will be teaching their usual classroom unit on force and motion. As part of the research study, the class will take *two science quizzes* (no more than 30 minutes each)—one quiz before and one after the unit. Your student's grade will not be affected in any way by this quiz, and participation is completely voluntary.

- 2009 California English Language Development Test (CELDT) in California, or the Arizona English Language Learner Assessment (AZELLA) in Arizona.
- 2009 math and 2010 science California Standards Test (CST) in California, or Arizona's Instrument to Measure Standards (AIMS) in Arizona.

This information helps us evaluate the advanced teacher-training program in terms of its effect on various achievement gaps (e.g., White-Black, White-Hispanic, girl-boy, etc.). An important aim of the program is to close these and other gaps, and to study the role of the teacher-training program in doing so. Students' background information allows us to do these critical analyses that will help improve teaching and learning for all students.

Confidentiality

Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies your child, or their teacher, school, or district to anyone outside the study team, except as required by law.

No names of students, teachers, schools, or districts will appear with any quiz, recording, transcript, report, or other publication. All student names will be stored in a locked cabinet or password-protected computer file, to which only the research staff will have access.

Risks

We see no risks to your child from our use of the quizzes and information from district records.

Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies your child or your district to anyone outside the study team, except as required by law.

What we're asking you to do

A consent form and a return envelope are included with this letter. We hope you will sign the form and agree to let us include your child's data.

Whom to contact

For more information about the research, or your and your child's rights as a research participant, please contact Joan Heller or Erica Heath.

The Research
Principal Investigator
Joan I. Heller, PhD
Director
Heller Research Associates
510-873-0800 ext. 1
jheller@edservices.org

Your Rights
Independent Review Board
Erica Heath, CIP, MBA
President
Independent Review Consulting
415-485-0717
ejheath@irb-irc.com

Thank you for considering our request.

Parent Consent Form

Force and Motion for Teaching *2009/10*

I have read the description of this research.

[Please check one box.]

I DO ☐ **I DO NOT** ☐ –consent to my child's science quiz results being used in this research study.

–consent to your obtaining my child's scores on the following tests for use in this research study:

- 2009 state standardized test in math (CST or AIMS)
- 2010 state standardized test in science (CST or AIMS)
- 2009 state English development test (CELDT or AZELLA)

(If your child did not take this test, check here: ___ Test was not taken.)

–consent to your getting basic information from my child's school records, such as age, sex, ethnic background, grade in school, and other school-related data, from school records for use in this research study.

Student name (please print) _____ Student's date of birth (mm/dd/yy) _____

Parent or guardian name (please print) _____ Parent or guardian signature _____

Today's date _____

[T Name • School Name • Address • City, State, Zip]

Using the enclosed envelope, please return this form ***within two weeks*** *to your child's science teacher.*

If you have questions, please contact: Joan I. Heller, Ph.D.
510-873-0800 ext. 1 • jheller@edservices.org

Appendix F. California content standards in physical science reporting clusters

THE MOTION REPORTING CLUSTER

The following six California content standards are included in the Grade 8 Motion reporting cluster and are represented in this booklet by four test questions. These questions represent only some ways in which these standards may be assessed on the California Grade 8 Science Standards Test.

CALIFORNIA CONTENT STANDARDS IN THIS REPORTING CLUSTER

Motion	
8PC1.	**The velocity of an object is the rate of change of its position. As a basis for understanding this concept:**
8PC1.a.	*Students know* position is defined in relation to some choice of a standard reference point and a set of reference directions.
8PC1.b.	*Students know* that average speed is the total distance traveled divided by the total time elapsed and that the speed of an object along the path traveled can vary.
8PC1.c.	*Students know* how to solve problems involving distance, time, and average speed.
8PC1.d.	*Students know* the velocity of an object must be described by specifying both the direction and the speed of the object.
8PC1.e.	*Students know* changes in velocity may be due to changes in speed, direction, or both.
8PC1.f.	*Students know* how to interpret graphs of position versus time and graphs of speed versus time for motion in a single direction.

THE FORCES, DENSITY AND BUOYANCY REPORTING CLUSTER

The following 11 California content standards are included in the Grade 8 Forces, Density and Buoyancy reporting cluster and are represented in this booklet by six test questions. These questions represent only some ways in which these standards may be assessed on the California Grade 8 Science Standards Test.

CALIFORNIA CONTENT STANDARDS IN THIS REPORTING CLUSTER

Forces	
8PC2.	**Unbalanced forces cause changes in velocity. As a basis for understanding this concept:**
8PC2.a.	*Students know* a force has both direction and magnitude.
8PC2.b.	*Students know* when an object is subject to two or more forces at once, the result is the cumulative effect of all the forces.
8PC2.c.	*Students know* when the forces on an object are balanced, the motion of the object does not change.
8PC2.d.	*Students know* how to identify separately the two or more forces that are acting on a single static object, including gravity, elastic forces due to tension or compression in matter, and friction.
8PC2.e.	*Students know* that when the forces on an object are unbalanced, the object will change its velocity (that is, it will speed up, slow down, or change direction).
8PC2.f.	*Students know* the greater the mass of an object, the more force is needed to achieve the same rate of change in motion.
8PC2.g.	*Students know* the role of gravity in forming and maintaining the shapes of planets, stars, and the solar system.
Density and Buoyancy	
8PC8.	**All objects experience a buoyant force when immersed in a fluid. As a basis for understanding this concept:**
8PC8.a.	*Students know* density is mass per unit volume.
8PC8.b.	*Students know* how to calculate the density of substances (regular and irregular solids and liquids) from measurements of mass and volume.
8PC8.c.	*Students know* the buoyant force on an object in a fluid is an upward force equal to the weight of the fluid the object has displaced.
8PC8.d.	*Students know* how to predict whether an object will float or sink.

Appendix G. Student data obtained from district administrative records

Information on students obtained from district administrative records included demographic and test score data (table G1).

Table G1. Student data obtained from district administrative records

Data	*Format or code*
Date of birth according to district records	Date
Sex	F = female M = male
Race/ethnicity[a]	White Black Hispanic Asian American Indian Pacific Islander Other More than one
2008/09 Grade 7 mathematics (California Standards Test or Arizona's Instrument to Measure Standards)	Three-digit scaled score
English language proficiency classification as of summer 2009	EO = English only IFEP = initially fluent English proficient ELL = English language learner RFEP = reclassified fluent English proficient
Fall 2009 total English, listening, and speaking scale scores (California English Language Development Test or Arizona English Language Learner Assessment)	Three-digit scaled score
2009/10 Grade 8 science (California Standards Test or Arizona Instrument to Measure Standards)	Three-digit scaled score
2009/10 Grade 8 physical science (California Standards Test only)	
Reporting cluster 1: Motion	Number correct (0–8)
Reporting cluster 2: Forces, Density, and Buoyancy	Number correct (0–13)

a. White includes European; Black includes African American; Hispanic includes Latino and other Spanish origin; Asian includes Chinese, Indian, Japanese, Korean, and Vietnamese; American Indian includes Alaska Native; and Pacific Islander includes Filipino, Guamanian or Chamorro, Native Hawaiian, Samoan, and other Pacific Islander.
Source: Author.

Appendix H. Survey items used to measure teacher confidence

A survey was conducted to measure teachers' confidence in their ability to teach force and motion. Table H1 presents the results.

Table H1. Survey items used to measure teacher confidence in ability to teach force and motion

22. Please indicate how confident you are teaching the following concepts (whether or not they are currently included in your curriculum). (1 = not at all confident, 2 = not very confident, 3 = somewhat confident, 4 = very confident)
22a. An object that is moving with constant speed can have a changing velocity.
22b. An object moving at a constant speed has no overall or net force acting on it.
22c. The acceleration of an object is directly proportional to its net force.
22d. An unbalanced net force can cause an object to speed up OR slow down, depending on its direction.
22e. The force of gravity pulls harder on heavier objects than light but makes them all free-fall with the same acceleration.
22f. Speeding up is different from going fast.
22g. Acceleration can be speeding up, slowing down, or changing direction.
22h. An object moving at a constant speed has no overall or net force acting on it.
22i. Friction is a force.
23. Please indicate how confident you are in your ability to conduct the following activities in class. (1 = not at all confident, 2 = not very confident, 3 = somewhat confident, 4 = very confident)
23b. Teach students to collect and carefully record data.
23e. Balance time for student hands-on activities, reading assignments, lectures, and solving problem sets.
23f. Teach students to identify evidence or data that support an explanation.
23g. Help students learn to provide a scientific explanation for something that has been observed.
23h. Foster discussions among students that help them learn science.
23i. Get students to use scientific terms accurately.
23j. Teach students to articulate clear and convincing reasons for their answers.
23k. Teach science to students who have limited (intermediate) English proficiency.
23l. Effectively initiate and guide sense-making discussion among students.
24. To what extent do you agree or disagree with each of the following statements? (1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree, 5 = not applicable)
24c. Weaknesses in my knowledge about force and motion limit how well I teach the unit.
24d. I am a good teacher of force and motion because I understand the content myself.
24e. I know how to use the district force and motion curriculum (for example, Full Option Science System [FOSS], Glencoe/McGraw-Hill).
24g. I am skilled at analyzing my students' work to understand their thinking about force and motion.
24h. I know how to question students to find out what they really do and do not understand about force and motion.

Source: Teacher survey instrument developed by author.

Appendix I. Course session video recording protocol

Hints for Videotaping

As part of this study, it is necessary to videotape every session. This guide offers helpful hints for making this a little easier. While you are not expected to produce broadcast quality video, we have included "tips" for how best to capture what happens during the professional development, along with practical advice about keeping track of this important data.

Before you begin—a few practical matters

Camera person. While it is not necessary, it may be helpful to designate someone in your group as the "cameraperson." This can be the co-facilitator, a participant, or someone else on site. This cameraperson does NOT need to closely attend to the camera, for example to zoom in and out, or swing the camera to follow conversations. However, it is important for someone to shift the camera when participants move from small-group to whole-group interactions, and change and label tapes as needed.

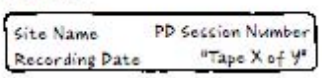
Equipment. Use the best quality equipment you have available. If you are unable to locate the necessary materials, please contact the Learning for Science Research staff.

Check that you have the following equipment:

- Mini DV Camera
- External Microphone
- Headphones
- Extension Cord
- Blank Mini DV tapes (2-3 per session)
  NOTE: Blank tapes are provided.

Handling the tapes. It may seem trivial to attend to the details of changing and labeling the tapes, yet this is critical and easy to overlook. The following tips may help:

Label the tapes. Before putting the tape in the camera, label the narrow end of each tape as follows:



Plan tape changes in advance. Plan to change the tapes during transitions in activity. Coordinate with your cameraperson or co-facilitator.

Don't worry about wasting tape. Change tapes during sensible transition times. It is better to have blank space at the beginning and end of the tape than to change tapes during key moments of the work or discussion.

Lock the tape. Immediately after removing the tape from the camera, lock the tape to prevent accidentally re-recording over your footage. To do this, look for the tab along one edge of the tape. Slide the tab so the plastic covers the opening.

Addressing anxiety. You may find some teachers are initially uncomfortable with the camera. This is often the case. However, once teachers get involved, they soon forget the camera is around. At the beginning of the first session, it may be helpful to acknowledge the camera and any feelings of discomfort people may have. Remind teachers that the videotapes are for research and educational purpose, and NOT to judge or report on individual teachers. As their consent forms indicated, the videotapes may be used in the following ways:

- Studied by research staff in this study
- Excerpted and incorporated in teacher education and research publications, along with materials used to promote those publications
- Shown to students in teacher education programs and at meetings of educators and educational researchers

Enhancing audio quality—helpful hints

Getting good audio is perhaps the most difficult aspect of videotaping in this setting. Flat echoing walls and multiple conversations happening simultaneously can make good sound retrieval a challenge. The following tips may help:

- Use an external microphone.
- Check the sound occasionally by listening through the headphones.
- Reduce sources of extraneous noise. If possible, consider closing doors and windows or turning off fans and air conditioners.

Enhancing video quality—helpful hints

Your goal is have sufficiently high video quality so it is possible to interpret what is being shown. The following tips may help:

Use a tripod. Set the tripod as high as possible to capture as much of the room as you can. You might try setting the tripod on a table.

Limit camera movement. Set the camera to capture the widest angle possible. If a cameraperson is monitoring the filming, avoid frequent zooming in and out, or panning across the room.

Increase the amount of light in the room. Turn on all of the lights and open the curtains.

Check legibility of board and charts. Check to see that the writing on the board, charts and easel pads is visible. Write with dark markers to improve on-camera visibility. Green pens are typically hard to see.

Avoid shooting into bright light. Face the camera away from windows or other sources of bright light. (Shooting with a bright background will cause participants to appear as silhouettes.)

Eliminate dead space. Look through the viewfinder to see what is being captured. Try not to capture too much empty space above the participants' heads.

Whole Group

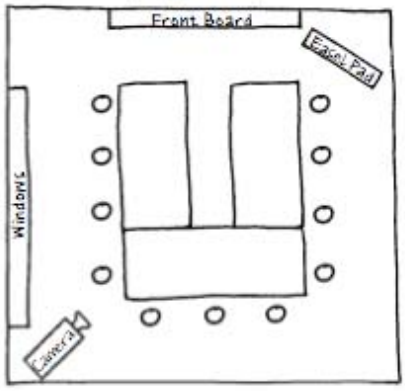
Following are recommendations for videotaping whole group activities, such as discussions, demonstrations, etc.

Audio Setup. By listening with headphones, test to see that the microphone is picking up the voices of participants from different locations in the room.

Camera Movement. In a large-group discussion, avoid trying to follow a conversation back and forth between different people. You will find that the camera always arrives late to the action. It is best to leave the camera set on a wide angle and only zoom in for a few seconds when it is necessary to read something on a chart or the board. Alternatively, after the session is over, you can shoot close-ups to capture what has been drawn or written down.

Camera Position. It is optimal to place the camera on a tripod off to the side in the back of the room (as shown below). Place it as high as possible, perhaps on a countertop or table. Your goals are to capture the facilitator, what is written at the front of the room, and the largest number of participants (preferably showing their faces).

Whole group camera position



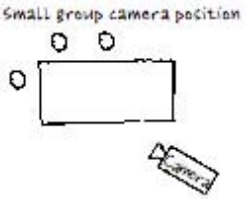
Small Group

Following are recommendations for videotaping small group activities, such as hands-on work and discussion in groups-of-three.

Audio Setup. By listening with headphones, test to see that the microphone is picking up the voices of participants in the small group over the din of other sounds in the room.

Camera Movement. Use a wide-angle setting in order to show all group members and leave room to capture any facilitators who might interact with the group.

Camera Position. It is optimal to place the camera on a tripod near one small group. Place it as high as possible, perhaps on a countertop or table, in order to look in and capture how teachers use the materials. If participants are working in groups of three, it helps if they are seated around the corner of a table (as shown below). Your goals are to capture all members of the small group, show how they are using hands-on and/or print materials, and any interactions with facilitators (preferably showing everyone's faces).



Appendix J. Course session attendance sheet

Attendance Sheet—Monday PD

Force and Motion for Teaching *2009/10*

Date: _____ Site: _____

Start time: _____ Facilitator: _____

End time: _____ Facilitator: _____

	Name	Time In AM	Time Out AM	Time In PM	Time Out PM	Initials
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						

Appendix K. Student test administration instructions for proctors

Instructions for Proctors

Force and Motion for Teaching *2009/10*

Now

☐ **Read these instructions now**, and familiarize yourself with the following tasks.

Administering Student Quiz 1

☐ **Find your materials.** You should have received from the teacher the materials you'll need:
- A blue envelope for this class, containing Student Quiz 1 and Quiz 2 Answer Sheets. *Do not detach Quiz 1 answer sheets from Quiz 2 Answer Sheets until directed to do so.*
- Student Quiz Booklets
- Pencils

☐ **Hand out the quiz booklets, answer sheets, and pencils.**
- Ask students to write their first and last name on the white label on Student Quiz 2 Answer Sheet.
- Ask students to separate the Quiz 1 and 2 Answer Sheets and hand in the Quiz 2 Answer Sheet. Put these answer sheets back in the blue envelope.
- Ask students to write today's date, class period, and their grade on the Student Quiz 1 Answer Sheet.

☐ **Read the following instructions to students:**

This is a quiz to find out how you think about force and motion before you begin learning about it in class. The quiz has some questions that are pretty easy, and some that are really hard, so you probably will not know the answers to some of them. If you are not sure of an answer, just make your best guess–there is no penalty for guessing.

If you do not know a word, raise your hand and I will read the question to you. However, I cannot help you answer the questions on the quiz.

Please do not write in the quiz booklet. This booklet will be used for other classes. Use only the answer sheet to record your answers.

You have until the end of this period to finish. [Tell students what to do if they finish early.]

☐ **Answer any student questions about the quiz.**
☐ **Monitor the class** during the quiz, to ensure that students are working independently.
☐ **Collect all answer sheets, quiz booklets, and pencils.**

After quizzes are handed back

☐ **Check each answer sheet** to make sure that ALL questions are answered. If necessary, return quizzes to students and encourage them to choose an answer for each question.
☐ **Return the completed answer sheets.** Place all of the Quiz 1 answer sheets into the enclosed white envelope labeled Completed Quiz 1 Answer Sheets. Seal the envelope, and put it in the blue envelope for this class.
☐ **Complete your Proctor Payment Form** after you have finished administering Quiz 1 to all classes. Place the form into the blue envelope for this class. This is an essential step as it is how we know what quizzes you proctored, and it is how you get paid.

Administering Student Quiz 2

- ☐ **Find your materials.** You should have received from the teacher the materials you'll need:
  - Blue envelope for this class
  - Student Quiz Booklets
  - Pencils
- ☐ **Hand out the quiz booklets and pencils.**
- ☐ **The teacher should hand out the Student Quiz 2 Answer Sheets**, ensuring that each student has the answer sheet with their name on it. If there is a student with no answer sheet, take one of the unused answer sheets *from this class's blue envelope* and have the student write his or her name on the white label on the Quiz 2 answer sheet.
- ☐ **Ask students to write today's date on the Student Quiz 2 Answer Sheet.**
- ☐ **Read the following instructions to students:**

  This is a quiz to find out how you think about force and motion. The quiz has some questions that are pretty easy, and some that are really hard, so you probably will not know the answers to some of them. If you are not sure of an answer, just make your best guess–there is no penalty for guessing.

  If you do not know a word, raise your hand and I will read the question to you. However, I cannot help you answer the questions on the quiz.

  Please do not write in the quiz booklet. This booklet will be used for other classes. Use only the answer sheet to record your answers.

  You have until the end of this period to finish. [Tell students what to do if they finish early.]

- ☐ **Answer any student questions about the quiz.**
- ☐ **Monitor the class** during the quiz, to ensure that students are working independently.
- ☐ **Collect all answer sheets, quiz booklets, and pencils.**

After quizzes are handed back

- ☐ **Check each answer sheet** to make sure that ALL questions are answered. If necessary, return quizzes to students and encourage them to choose an answer for each question.
- ☐ **Return the completed answer sheets.** Place all of the Quiz 2 answer sheets into the blue envelope for this class. (There is no white envelope for Quiz 2.)
- ☐ **Complete your Proctor Payment Form** after you have finished administering Quiz 2 to all classes. Place the form into the blue envelope for this class. This is an essential step as it is how we know what quizzes you proctored, and it is how you get paid.

Questions? Please contact Cara Peterman
at hra@edservices.org or 510-873-0800 ext. 4.

Appendix L. Teacher test administration instructions for site coordinators

Instructions for Site Coordinators

The following is a summary of the data collection during Teacher Meeting 1, including specific instructions for you and the participants.

Introduce the data collection

- ❑ Please tell teachers that will be asked to complete three research instruments today for the study—a **consent form**, **survey**, and **science quiz**. This should take no more than an hour altogether.
- ❑ Talk about the importance of the research components and the general plan for collecting data, as outlined in the consent form. The following may be useful talking points:

  We are filling out these forms because, as you know, this course is part of a large-scale state-wide research study, funded by the U.S. Department of Education. The research examines ways to improve student achievement by strengthening the training that teachers receive in science.

  As the teachers involved in this research, you have a crucial role in determining the value of the results. The potential of this study depends upon the collection of data from you and your students, so at several points over the next two years, you will be asked to help with the course evaluation.

  Not only are you likely to benefit from this course, but other teachers and their students across the state stand to gain from what is learned through your participation.

  Please help protect the study's integrity by NOT discussing the details of this course with other teachers.

Remind teachers when data will be collected

During Teacher Meeting 1 and Teacher Meeting 2. Each of you will complete a 30-minute survey and a 30-minute science quiz today and again during Teacher Meeting 2.

Before and after your classroom force and motion unit. You will also each give your students a science quiz before and after your force and motion unit, and provide some background information about your classroom. Each of you will get a packet in the mail explaining everything you need to know about collecting student data.

Several of you will provide additional data through interviews and classroom visits by researchers as a way to capture the complexity of teaching that is not possible in a pencil-and-paper survey.

Note for teachers: Please complete all data collection on time. If this is not possible for any reason, let the research staff know what to expect by contacting:

Cara Peterman cpeterman@edservices.org

Collect the data

- ❑ **Hand out and collect a signed Teacher Consent Form from each participant.** Read these instructions to teachers.

  Please read and sign to indicate that you understand what you will be doing as part of this research project.

- ❑ **Administer the Science Teaching Survey.** Read these instructions to teachers.

  This survey asks about your beliefs and practices related to teaching force and motion to students. Start by completing the information and ID numbers on the cover sheet and first page of the survey.

  In order to keep your data anonymous and confidential, the cover sheets with your names will be removed upon receipt by the research staff, leaving only the ID numbers on the first page of the survey. The cover sheets will be stored in a locked cabinet, separate from the completed surveys.

  If you are not sure how to interpret a question, do the best you can, but also write a note to the researchers in the margin so they are alerted to the problem.

- ❑ **Administer the Teacher Quiz.** Read these instructions to teachers.

  We have allotted 30 minutes for you to fill out the survey. Start by completing the information and ID numbers on the cover sheet and first page of the survey.

  The quiz is designed to include questions with a wide range of difficulty, and we expect you to encounter items for which you may not know the answers. This is especially true before you have taken the course! If you are not sure of an answer, please make your best guess—there is no penalty for guessing. And keep in mind that you get another chance to answer questions like these after the course!

  As with the survey, if you are not sure how to interpret a question, do the best you can, but also write a note to the researchers in the margin so they are alerted to the problem.

Check and return the written data

- ❑ **Double-check the surveys and quizzes** to make sure ALL pages, including the cover page and next page, are completely filled out. If necessary, return surveys or quizzes to the teachers and encourage them to answer all questions rather than leave any blank.

- ❑ **Put all copies of the consent forms, surveys, and quizzes into the return envelope provided** to be mailed to Heller Research Associates. For the validity of the evaluation, please do not keep copies of these documents. Return ALL unused surveys or quizzes in the HRA envelope.

Appendix M. Baseline equivalence of teacher demographics in intervention and control group samples

No statistically significant differences in teacher demographic characteristics were found between the intervention and control groups in the full recruited sample of teachers (table M1), the retained teacher sample (table M2), or the sample that was not retained (table M3). About 60 percent of the retained teacher sample were women, almost 75 percent were White, and almost 90 percent were native English speakers.

Table M1. Teacher demographic information for full teacher sample, by experimental condition

	Intervention group		*Control group*		
Characteristic	*Number*	*Percent*[a]	*Number*	*Percent*[a]	p-*value* [b]
Sex					.54
Female	58	64.4	54	59.3	
Male	32	35.6	37	40.7	
Race/ethnicity					.58
White	59	65.6	62	68.1	
Black	3	3.3	4	4.4	
Hispanic	13	14.4	10	11.0	
More than one race	10	11.1	6	6.6	
Other or unknown	5	5.5	9	9.9	
English language status					.51
Entered school speaking little or no English	5	5.6	10	11.0	
Entered school speaking enough English to participate in some classroom interactions	0	0	0	0	
Entered school speaking enough English to participate in most classroom interactions	3	3.3	3	3.3	
Nonnative English speaker but entered school fully English proficient	#	#	4	4.4	
Native English speaker	79	87.8	73	80.2	
Unknown	#	#	#	#	
Home or primary language					.63
English	84	93.3	82	90.1	
Spanish	#	#	3	3.3	
Other or unknown	4	4.5	6	6.6	

Note: $n = 90$ for intervention group, $n = 92$ for control group. White includes European; Black includes African American; Hispanic includes Latino and other Spanish origin; Asian includes Chinese, Indian, Japanese, Korean, and Vietnamese; American Indian includes Alaska Native; and Pacific Islander includes Filipino, Guamanian or Chamorro, Native Hawaiian, Samoan, and other Pacific Islander.

# indicates values were suppressed to reduce disclosure risk.

a. Computed based on valid (non-missing) data. Components may not sum to 100 because of rounding.

b. Two-tailed Fisher's exact test for equality of proportion between intervention and control group teachers.

Source: Author's analysis of primary data collected for the study.

Table M2. Teacher demographic information for retained teacher sample, by experimental condition

	Intervention group		*Control group*		
Characteristic	*Number*	*Percent*[a]	*Number*	*Percent*[a]	*p*[b]
Sex					
Female	42	60.9	36	56.3	.60
Male	27	39.1	28	43.8	
Race/ethnicity					
White	47	68.1	49	76.6	.51
Hispanic	11	15.9	5	7.8	
More than one race	8	11.6	5	7.8	
Other or unknown	3	4.2	5	7.9	
English language status					
Entered school speaking little or no English	4	5.8	3	4.8	.33
Native English speaker	62	89.9	54	85.7	
Other	3	4.2	6	9.6	
Home or primary language					
English	65	94.3	61	95.3	.85
Other or unknown	4	5.6	3	4.7	

Note: $n = 69$ for intervention group, $n = 64$ for control group. White includes European; Black includes African American; Hispanic includes Latino and other Spanish origin; Asian includes Chinese, Indian, Japanese, Korean, and Vietnamese; American Indian includes Alaska Native; and Pacific Islander includes Filipino, Guamanian or Chamorro, Native Hawaiian, Samoan, and other Pacific Islander.

a. Computed based on valid (non-missing) data. Components may not sum to 100 because of rounding.

b. Two-tailed Fisher's exact test for equality of proportion between intervention and control group teachers.

Source: Author's analysis of primary data collected for the study.

Table M3. Teacher demographic information for not retained teacher sample, by experimental condition

	Intervention group		*Control group*		
Characteristic	*Number*	*Percent*[a]	*Number*	*Percent*[a]	*p*[b]
Sex					
Female	16	76.2	18	66.7	.54
Male	5	23.8	9	33.3	
Race/ethnicity					
White	12	57.1	13	48.1	.61
Black	#	#	3	11.1	
Hispanic	#	#	5	18.5	
Asian	#	#	3	11.1	
Other or unknown	4	19.0	3	11.1	
English language status					
Native English speaker	17	81.0	19	70.4	
Entered school speaking little or no English	#	#	7	25.9	.11
Other	3	14.3	1	3.7	
Home or primary language					
English	19	90.5	21	77.8	.59
Other	#	#	6	22.2	

Note: $n = 21$ for intervention group, $n = 27$ for control group. White includes European; Black includes African American; Hispanic includes Latino and other Spanish origin; Asian includes Chinese, Indian, Japanese, Korean, and Vietnamese; American Indian includes Alaska Native; and Pacific Islander includes Filipino, Guamanian or Chamorro, Native Hawaiian, Samoan, and other Pacific Islander.
# indicates values were suppressed to reduce disclosure risk.
a. Computed based on valid (non-missing) data. Components may not sum to 100 because of rounding.
b. Two-tailed Fisher's exact test for equality of proportion between intervention and control group teachers.
Source: Author's analysis of primary data collected for the study.

Analyses to determine whether the groups were equivalent at baseline with respect to teacher education, training, and experience indicated that no more differences between intervention and control groups were detected within the recruited (table M4), retained (table M5), or not retained (table M6) samples than would have been expected based on chance. The only comparison for which a significant difference was detected was for retained teachers in the number of semesters of postsecondary classes taken in science (table M4): control group teachers took more such classes than intervention group teachers. Most participants were experienced teachers, averaging about 11 years of teaching experience, 9 years of experience teaching science, 6 years of experience teaching force and motion, and more than 8 years of experience teaching English language learners for all samples.

Table M4. Teacher education, training, and experience at baseline for full recruited teacher sample, by experimental condition

(percent of sample)

Measure	*Intervention group*	*Control group*	*Difference*	p-*value*[a]
Teacher education				
Type of teaching certification				
Permanent or standard	76.7	75.8	0.9	≥.99
Cross-cultural or language development (for example, Crosscultural, Language, and Academic Development [CLAD])	33.3	38.5	–5.2	.54
Subject area/level of teaching certification				
Science	72.2	70.3	1.9	.87
Multiple subject	37.8	31.9	5.9	.44
Bachelor's degree in science	58.9	62.6	–3.7	.65
Number of semesters of postsecondary classes taken				
Science				.23
0–2	26.7	27.5		
3–4	40.0	28.6		
5 or more	33.3	44.0		
Methods of teaching science				.33
0–2	82.2	87.9		
3–4	13.3	6.6		
5 or more	4.4	5.5		
Teaching English language learners				.68
0–2	85.6	87.9		
3 or more	14.6	12.1		
Hours of professional development in last three years focused on force and motion				
Mean	17.2	13.8	3.3	.14
Standard deviation	27.7	30.3		
n	89	91		
Teaching experience				
Years as a teacher				
Mean	11.4	11.1	0.3	.63
Standard deviation	8.4	9.1		
n	90	91		
Years teaching science				
Mean	9.0	9.2	–0.2	.85
Standard deviation	6.9	8.0		
n	90	91		
Years teaching force and motion				

Measure	*Intervention group*	*Control group*	*Difference*	p-*value*[a]
Mean	6.0	6.6	–0.6	.79
Standard deviation	4.9	6.3		
n	90	91		
Years teaching English language learners				
Mean	9.6	8.5	1.0	.24
Standard deviation	6.8	6.8		
n	90	91		

a. *p*-value for quantitative data determined through Monte Carlo estimation of exact Wilcoxon rank sum test. *P*-value for categorical data determined through two-tailed Fisher's exact test.
Source: Author's analysis of primary data collected for the study.

Table M5. Teacher education, training, and experience at baseline for retained teacher sample, by experimental condition

Measure	*Intervention group*	*Control group*	*Difference*	p-*value*[a]
Teacher education				
Type of teaching certification				
Permanent or standard	84.1	79.7	4.4	.65
Cross-cultural or language development (for example, Crosscultural, Language, and Academic Development [CLAD])	27.5	32.8	–5.3	.57
Subject area/level of teaching certification				
Science	71.0	71.9	-0.9	≥.99
Multiple subject	34.8	28.1	6.7	.46
Bachelor's degree in science	59.4	67.2	–7.8	.37
Number of semesters of postsecondary classes taken				
Science				
0–2	26.1	28.1	-2.0	.03
3–4	42.0	21.9	20.1*	
5 or more	31.9	50.0	-18.1	
Methods of teaching science				
0–2	84.1	85.9		.71
3 or more	16.0	14.1		
Teaching English language learners				
0–2	85.5	89.1		.79
3 or more	14.5	11.0		
Hours of professional development in last three years focused on force and motion				
Mean	16.5	15.0	1.5	.27
Standard deviation	26.4	33.8		

Measure	*Intervention group*	*Control group*	*Difference*	p-*value*[a]
n	68	64		
Teaching experience				
Years as a teacher				
Mean	11.6	11.2	0.4	.79
Standard deviation	8.6	8.5		
n	69	64		
Years teaching science				
Mean	9.3	9.3	–0.1	.93
Standard deviation	7.0	7.5		
n	69	64		
Years teaching force and motion				
Mean	5.8	6.9	–1.2	.43
Standard deviation	4.6	6.2		
n	69	64		
Years teaching English language learners				
Mean	10.1	8.7	1.4	.22
Standard deviation	7.0	6.5		
n	69	64		

*Significantly different from zero at the .05 level, two-tailed test.
a. *p*-value for quantitative data determined through Monte Carlo estimation of exact Wilcoxon rank sum test. *P*-value for categorical data determined through two-tailed Fisher's exact test.
Source: Author's analysis of primary data collected for the study.

Table M6. Teacher education, training, and experience at baseline for not retained teacher sample, by experimental condition

Measure	*Intervention group*	*Control group*	*Difference*	*p*-value[a]
Teacher education				
Type of teaching certification				
Permanent or standard	52.4	66.7	–14.3	.38
Cross-cultural or language development (for example, Crosscultural, Language, and Academic Development [CLAD])	52.4	51.9	0.5	≥.99
Subject area/level of teaching certification				
Science	76.2	66.7	9.5	.54
Multiple subject	47.6	40.7	6.9	.77
Bachelor's degree in science	57.1	51.9	5.2	.78
Number of semesters of postsecondary classes taken				
Science				
0–2	28.6	25.9	0.7	.76
3–4	33.3	44.4	-11.1	
5 or more	38.1	29.6	8.5	
Methods of teaching science				

Measure	*Intervention group*	*Control group*	*Difference*	*p*-value[a]
0–2	76.2	92.6	–16.4	.25
3–4	19.1	3.7	15.4	
5 or more	4.8	3.7	1.1	
Teaching English language learners				
0–2	85.7	85.2	0.5	.86
3–4	14.3	11.1	3.2	
5 or more	0.0	3.7	–3.7	
Hours of professional development in last three years focused on force and motion				
Mean	19.2	11.0	8.2	.32
Standard deviation	32.0	19.7		
n	21	27		
Teaching experience				
Years as a teacher				
Mean	10.8	10.9	–0.1	.77
Standard deviation	7.9	10.5		
n	21	27		
Years teaching science				
Mean	8.2	8.7	–0.5	.85
Standard deviation	6.8	9.1		
n	21	27		
Years teaching force and motion				
Mean	6.6	5.7	0.9	.76
Standard deviation	6.2	6.6		
n	21	27		
Years teaching English language learners				
Mean	7.8	8.2	–0.4	.96
Standard deviation	6.1	7.7		
n	21	27		

a. *p*-value for quantitative data determined through Monte Carlo estimation of exact Wilcoxon rank sum test. *P*-value for categorical data determined through two-tailed Fisher's exact test.
Source: Author's analysis of primary data collected for the study.

Appendix N. Class selection worksheet

The student sample was determined at the class level through random selection of two grade 8 physical science classes per teacher. All physical science classes were considered eligible except those that included only special education students. The classes were selected using a class selection worksheet developed for this purpose (table N1). The worksheet led teachers through a process of identifying and numbering their eligible grade 8 science classes and then determining in which of those classes to collect data. The key to randomizing each teacher's classes was a random number selection table that was unique to each teacher. These tables were created using randomly generated numbers and then merged into the worksheets, so that no two teachers received the same class selection criteria. If a teacher taught only one or two eligible class sections, student data were collected from those sections. For teachers who taught three grade 8 science classes, the table provided two random numbers between one and three; for teachers who taught four eligible classes, the table provided two random numbers between one and four; and so forth. The table was included in each teacher's student data packet.

Table N1. Example of personal random number selection table included in each teacher's class selection worksheet

If your number of eligible classes is	*Then the classes you collect data in are*
1	Your one eligible science class.
2	Both of your eligible science classes.
3	Your first and third eligible science classes.
4	Your second and third eligible science classes.
5	Your fourth and fifth eligible science classes.
6	Your first and second eligible science classes.

Note: All grade 8 physical science classes were eligible except those comprising only special education students.
Source: Class selection worksheet developed by author.

Appendix O. Sensitivity analysis for nesting of students within teachers or classes within teachers

More than 90 percent of teachers submitted two class sets of student data. To determine whether it was necessary to nest students within classes within teachers in the student models or whether it was sufficient to nest students within teachers, the study team examined the sensitivity of impact estimates to these alternatives (table O1). There were no differences between impact estimates in models in which students were nested only within teachers and models in which students were nested within both teachers and classes within teachers.

Table O1. Sensitivity of student impact estimates to alternative model specification: nesting of students within teachers versus nesting of students within classes within teachers

	Adjusted mean (standard deviation)					*Unweighted*	
Covariate	*Intervention group*	*Control group*	*Difference (standard error)*	p-*value*	*Confidence interval*	*Effect size*	*Student (teacher) sample size*
ATLAST Test of Force and Motion							
Students within teacher	52.4 (19.8)	50.3 (19.3)	2.1 (1.0)	.04	0.4–3.7	0.11	5,130 (127)
Students within class within teacher	52.3 (19.8)	50.2 (19.3)	2.0 (1.0)	.04	0.4–3.7	0.11	5,130 (127)
California Standards Test physical science reporting clusters							
Students within teacher	71.0 (19.4)	70.4 (19.4)	0.5 (1.1)	.62	–1.3 to 2.4	0.03	3,768 (96)
Students within class within teacher	70.9 (19.4)	70.4 (19.4)	0.4 (1.1)	.69	–1.4 to 2.3	0.02	3,768 (96)

ATLAST is Assessing Teacher Learning About Science Teaching.
Note: All models were estimated with student sample with valid non-missing posttest data. Data were adjusted using multilevel regression models to account for differences in baseline and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable. Model used full set of covariates:
Student demographic characteristics: sex (male, female); English language learner status (English language learner, fluent English proficient); and race/ethnicity (White, Hispanic, Black, Asian, Other).
Student pretest measure of outcome variable (ATLAST Test of Force and Motion pretest; standardized grade 7 mathematics scale scores from 2008/09 in lieu of California Standards Test pretest).
Teacher (random intercept).
Teacher pretest measure of content knowledge (ATLAST Test of Force and Motion pretest of force and motion).
Teacher teaching experience, based on ordinal five-level scale: beginning (0–2 years), high beginning (3–4 years), middle (5–7 years), high middle (8–10 years), and veteran (11 or more years).
Teacher undergraduate degree (science, not science).
Treatment group (intervention, control).
Site-by-treatment interaction.
Teacher randomization stratum.
Missing-value indicators.
Source: Author's analysis of primary data collected for the study.

Appendix P. Impact estimation methods

The primary student-level model is a hierarchical linear model for a continuous outcome:

$$Post_{ijk} = \mu + \beta_{pre}\,\mathrm{Pre}_{ijk} + \sum_{r=1}^{m}\beta_M^r M_k^r + \beta_{Tx}\,\mathrm{Tx}_{jk} + \sum_{r=1}^{s}\beta_{Tx}^r\left(S_k^r Tx_{jk}\right) + Sex_{ijk} + \beta_{EL}EL_{ijk} + \beta_R R_{ijk} +$$

$$\sum_{r=1}^{5}\beta_E^r TeachExp_{jk}^r + \beta_{know}TeachKnow_{jk} + \beta_{TS}TeachSex_{jk} + \beta_{bach}Bach_{jk} + \beta_{Mbach}MissBach_{jk} + \tau_{jk} + \varepsilon_{ijk} \quad \text{(P1)}$$

where subscript *i* denotes the student stratum, *j* denotes the teacher stratum, and *k* denotes the randomization stratum, and all variables other than the pretests are dummy variables (table P1).

Table P1. Variables included in hierarchical linear models for student-level outcomes

Variable	*Term*	*Description*
Outcome variable	Post	Posttest measure of outcome variable.
Pretest	Pre	Baseline or pretest measure of outcome variable (for teachers, ATLAST and baseline confidence in teaching force and motion from teacher survey ratings; for students, baseline for 2009/10 California Standards Test physical science reporting clusters is standardized grade 7 mathematics scale scores from 2008/09).
Teacher randomization stratum	M_k^r	Dichotomous variables for being in stratum *r*, *r = 1, ..., M*, where *M* represents the number of blocks. The coefficients to these variables are the estimated differences between mean outcome for that stratum and the mean for all blocks. The sum of the coefficients was constrained to sum to zero.
Treatment group of teacher	Tx	Dichotomous variable indicating whether the student's teacher was assigned to the intervention condition.
Site-by-treatment interaction	$S_k^r Tx_{jk}$	Dichotomous variable for whether a given teacher j_k was both treated *and* in site *r* = 1*, ..., S*, with *S* being the number of sites. The sum of the coefficients was constrained to sum to zero.
Student sex	Sex_{ijk},	Dichotomous variable (1 indicates female, 0 indicates male).
Student English language learner status	EL_{ijk},	Dichotomous variable (1 indicates English language learner, 0 indicates fluent English proficient).
Student race/ethnicity	R_{ijk},	Set of dichotomous variables for White, Hispanic, Black, Asian, and Other.

Variable	*Term*	*Description*
Teacher pretest	*TeachKnow*	Baseline or pretest measure of teacher's content knowledge (ATLAST test score).
Teacher sex	*TeachSex*	Dichotomous variable (1 indicates female, 0 indicates male).
Teacher Bachelor's degree	*Bach*	Dichotomous variable (1 indicates undergraduate degree in science, 0 indicates no undergraduate degree in science).
Teaching experience	*TeachExp*	Control variable for years of teaching experience, based on ordinal, five-level scale: beginning (0–2 years), high beginning (3–4 years), middle (5–7 years), high middle (8–10 years), and veteran (11 or more years).
Missing-value indicators	*MissX*	Variable for measure *X* is missing. One set of indicators for each measure with any missing values.
Teacher	τ_{jk}	Random intercept for teacher, assumed to be normally distributed with zero mean and variance to be estimated from data.
Error	ε	Error term for individual students.

ATLAST is Assessing Teacher Learning About Science Teaching.
Source: Author.

To assess the overall impact of the intervention on all students and on English language learner students, model 1 was estimated on both samples. For each population, this model was estimated twice, once for the results of the ATLAST Test of Force and Motion and once for the results of the 2009/10 California Standards Test physical science reporting clusters. The random effect (intercept) of teacher is captured by τ_{jk}, which accounts for the positive intraclass correlations in the data.

The primary model for teacher-level outcomes is:

$$Post_{jk} = \mu + \beta_{pre}\,\mathrm{Pre}_{jk} + \sum_{r=1}^{m}\beta_M^r M_k^r + \beta_{Tx}\,\mathrm{Tx}_{jk} + \sum_{r=1}^{s}\beta_{Tx}^r\left(S_k^r Tx_{jk}\right) + \beta_{Abl}Abl_{jk}$$
$$+\beta_{PC}PreConf_{jk} + \beta_{MPC}MissPreConf_{jk} + \sum_{r=1}^{5}\beta_E^r TeachExp_{jk}^r +$$
$$\sum_{r=1}^{5}\beta_{TS}TeachSex_{jk} + \beta_{bach}Bach_{jk} + \beta_{Mbach}MissBach_{jk} + \varepsilon_{jk} \qquad \text{(P2)}$$

where subscripts *i* denotes student, *j* denotes teacher, and *k* denotes randomization stratum, and all variables other than the pretests are dummy variables (see description of variables in teacher-level models in table P2).

Table P2. Variables included in hierarchical linear models for teacher-level outcomes

Variable	*Term*	*Description*
Outcome variable	Post	Posttest measure of outcome variable.
Pretest	Pre	Baseline or pretest measure of outcome variable (ATLAST Test of Force and Motion for teachers and baseline confidence in teaching force and motion from teacher survey ratings).
Teacher randomization stratum	M_k^r	Dichotomous variables for being in stratum *r*, *r = 1, …, M*, where *M* represents the number of blocks. The coefficients to these variables are the estimated differences between mean outcome for that stratum and the mean for all blocks. The sum of the coefficients was constrained to sum to zero.
Treatment group of teacher	*Tx*	Dichotomous variable indicating whether the teacher was assigned to the intervention condition (T indicates treatment group; C indicates control group).
Site-by-treatment interaction	$S_k^r Tx_{jk}$	Dichotomous variable for whether a given teacher *jk* was both treated *and* in site $r = 1, \ldots, S$, with *S* being the number of sites. The sum of the coefficients was constrained to sum to zero.
Teacher sex	*TeachSex*	Dichotomous variable (1 indicates female, 0 indicates male).
Teacher Bachelor's degree in science	*Bach*	Dichotomous variable (1 indicates undergraduate degree in science, 0 indicates no undergraduate degree in science)
Teaching experience	*TeachExp*	Control variable for years of teaching experience, based on five-level scale: beginning (0–2 years), high beginning (3–4 years), middle (5–7 years), high middle (8–10 years), and veteran (11 or more years).
Teacher initial confidence	*PreConf*	Measured by teacher surveys (included as covariate in analysis of ATLAST test scores).
Missing-value indicators	*MissX*	Indicates whether measure *X* is missing. One set of indicators for each measure with at least one missing value.
Student academic ability	*Abl*	Teacher-aggregated student grade 7 scores on 2008/09 standardized test in mathematics.
Error	ε	Error term for individual teachers.

ATLAST is Assessing Teacher Learning About Science Teaching.
Source: Author's summary.

The coefficient of primary interest is β_{Tx}, the treatment effect. Fixed effects included in both student- and teacher-level models include baseline (pretest) measure of each outcome variable, randomization stratum of the teacher, site-by-treatment interaction, experimental condition of the teacher, and a teacher covariate for years of teaching experience.

In models of both student- and teacher-level outcomes, the coefficients for stratum, site-by-treatment, and teaching experience terms are each constrained to sum to zero. Because the sum of the coefficients to the site-by-treatment interaction terms is constrained to be zero, this impact estimate is the simple unweighted average of the impacts estimated for all six sites; the standard error is the variance of this parameter. This constraint on the interaction is equivalent to estimating six site treatment effects and computing the pooled estimate and variance from a simple mean contrast of those six estimates (Dynarski et al. 2004).

Appendix Q. Missing item–level data

Table Q1. Missing item–level data for student and teacher outcome measures

Outcome measure	Total sample		Intervention group		Control group		Percentage difference between groups	p-value[a]
	Number	*Percent*	*Number*	*Percent*	*Number*	*Percent*		
ATLAST Test of Force and Motion for Students								
Pretest								
Students without missing items	4,901	92.4	2,477	91.9	2,424	92.9	0.5	.92
Students missing 1–3 items	226	4.2	115	4.3	111	4.2		
Students missing 4–26 items	24	0.4	19	0.7	5	0.2		
Students missing all 27 items	154	2.9	84	3.1	70	2.7		
Students with any missing items	404	7.5	218	8.1	186	7.1		
Posttest								
Students without missing items	4,905	92.5	2,487	92.3	2,418	92.6	0.3	.91
Students missing 1–3 items	195	3.7	99	3.7	96	3.7		
Students missing 4–26 items	21	0.4	5	0.2	16	0.6		
Students missing all 27 items	184	3.5	104	3.9	80	3.1		
Students with any missing items	400	7.6	208	7.7	192	7.4		
ATLAST Test of Force and Motion for Teachers								
Preintervention								
Number of missing items (range)	0–1		0–1		0–1			
Number of teachers with missing items	6	4.5	4	5.8	#	#	2.7	.85
Postinstruction								
Number of missing items (range)	0–1		0–1		0–1			
Number of teachers with missing items	8	0.1	#	#	#	#	6.5	.73
Teacher confidence in ability to teach force and motion (23-item scale)								
Preintervention								
Number of missing items (range)	0–5		0–2		0–5			

Outcome measure	Total sample		Intervention group		Control group		Percentage difference between groups	p-value[a]
	Number	*Percent*	*Number*	*Percent*	*Number*	*Percent*		
Number of teachers with missing items	9	5.0	5	5.6	4	4.4	1.2	.89
Postinstruction								
Number of missing items (range)	0–1		0–1		0–1			
Number of teachers with missing items	6	4.5	#	#	#	#	6.4	.69

Note. ATLAST is Assessing Teacher Learning About Science Teaching.
# indicates data values suppressed to reduce disclosure risk.
a. Test for equality of proportion between intervention and control group teachers.
Source: Author's analysis of primary data collected for the study.

Appendix R. Schedule and content goals of Making Sense of SCIENCE™ professional development course on force and motion

The 24-hour Making Sense of SCIENCE™ teacher course on force and motion was taught over five days (table R1).

Table R1. Schedule for five-day Making Sense of SCIENCE™ course on force and motion

Day	*Morning (3 hours)*	*Afternoon (3 hours)*
1	Session 1, Part 1 Science investigation	Session 1, Part 2 Literacy analysis Case discussion Lesson planning
2	Session 2 Science investigation Literacy analysis	Session 3, Part 1 Science investigation Literacy analysis
3	Session 3, Part 2 Case discussion Lesson planning	No session
4	Session 4, Part 1 Science investigation Literacy analysis	Session 4, Part 2 Case discussion Lesson planning
5	Session 5 Science investigation	No session

Source: Draft schedule developed for and subsequently published in final form in: Daehler, K. R., Shinohara, M., and Folsom, J. (2011). *Making Sense of SCIENCE™: Force and motion for teachers of grades 6–8*. San Francisco, CA: WestEd.

Each course session addressed particular science content and literacy goals (table R2).

Table R2. Content of Making Sense of SCIENCE™ course on force and motion, by session

Session/topic	Goals
1: Motion	• Learn what this professional development course is about and how it is organized. • Interpret and represent motion using numbers, difference tables, number lines, calculations, illustrations, and graphs. • Differentiate between negative velocity and negative position. • Explore common ideas that students and teachers have about velocity and speed. • Consider how best to help students understand velocity. • Recognize the complexities and demands of science reading.
2: Changes in motion	• Understand acceleration. • Differentiate between negative acceleration and slowing down. • Explore common ideas that students and teachers have about acceleration and speed. • Consider how best to help students navigate the various languages and representations of acceleration, while steering clear of overly complex examples. • Examine strategies that support reading in science.
3: Acceleration and force	• Investigate how acceleration is affected by force. • Interpret events involving balanced and unbalanced forces. • Explore common ideas that students and teachers have about how things move and how forces act over time. • Figure out ways to help students think about the effects of forces over time (for example, constant versus impulse forces) and understand the role of initial motion. • Identify the challenges and supports of reading data.
4: Force	• Understand force as an interaction between objects. • Use arrows to represent forces and combinations of forces. • Recognize that an object slowing down due to friction is an example of a net force acting opposite the direction of motion. • Explore common ideas that students and teachers have about forces, especially friction. • Evaluate the utility of including "interaction" in the definition of force.
5: Acceleration and mass	• Understand how acceleration is affected by mass (and force). • Differentiate between mass and weight. • Explain how and why things fall the way they do on Earth. • Develop a one-year plan for teaching students to become better readers of science. • Reflect on and celebrate what individuals have learned about science, literacy, and the practice of teaching.

Source: Draft content developed for and subsequently published in final form in: Daehler, K. R., Shinohara, M., and Folsom, J. (2011). *Making Sense of SCIENCE™: Force and motion for teachers of grades 6–8.*. San Francisco, CA: WestEd.

Appendix S. Sensitivity analyses based on different models and analytic samples

To examine the robustness of the findings, the study team determined the sensitivity of findings to models estimated with different combinations of covariates and different analytic samples. Because teachers were randomly assigned to the intervention condition, the inclusion of covariates in the impact analysis model should theoretically have consequences only for the precision of the impact estimate, not for the point estimate itself. Changes in point estimates could arise from the inclusion of different sets of covariates because of baseline differences in characteristics across intervention and control groups. Differences in baseline characteristics, in turn, could reflect chance differences between groups at randomization or selective attrition after randomization.

Student outcomes

Impact analyses estimated primary student outcomes based on regression models that included different combinations of covariates (table S1) and analytic samples (table S2).

Influence of student-level covariates

Covariates were varied in three regression models:

- *Basic model*: Included no covariates beyond blocks dummy indicators, site × treatment interaction, and treatment condition.
- *Basic plus pretest model*: Included the variables in the basic model plus baseline test score and an indicator variable for missing data on the baseline student measure.
- *All covariates*: Included all of the above terms plus the student-level and teacher-level covariates described in chapter 2 and indicator variables for missing data on each applicable covariate.

All models were estimated for the student sample with valid non-missing posttests (n = 5, 130). Controlling for covariates did not significantly change estimates of impact on student outcomes. Estimates of impact on student scores on the ATLAST Test of Force and Motion or the California Standards Test physical science reporting clusters were not significant and varied little when covariates were included in the models.

Influence of analytic student sample

Treatment effects were estimated for three models involving different subsets of the student data:

- *Complete cases*: Student sample with valid non-missing pretest and posttest and complete data for all covariates (n = 4,612).
- *Pretest and posttest*: Student sample with valid non-missing pretest and posttest and missing covariate values replaced with the average of non-missing values (n = 4,967).
- *Posttest*: Student sample with valid, non-missing posttest and missing pretest and covariate values replaced with average of non-missing values (n = 5,130).

All models were estimated with the full set of covariates. Estimates of impacts for different analytic samples did not significantly change the student outcomes, which were not significant and varied little for different analytic samples.

Table S1. Sensitivity of student impact estimates to alternative model specifications

Measure/model	*Adjusted mean*			*Unadjusted p-value*	*Statistical significance after correction*[a]	*Effect size*	*Student sample size*
	Intervention Group (standard deviation)	*Control group (standard deviation)*	*Difference (standard error)*				
ATLAST Test of Force and Motion for students (percent correct)							
Basic model[b]	52.3 (19.8)	49.9 (19.3)	2.4 (1.5)	.10	No	0.13	5,130 (127)
Basic model plus pretest[c]	52.4 (19.8)	50.0 (19.3)	2.4 (1.1)	.03	No	0.12	5,130 (127)
All covariates[d]	52.4 (19.8)	50.3 (19.3)	2.1 (1.0)	.04	No	0.11	5,130 (127)
California Standards Test physical science reporting clusters							
Basic model[b]	71.0 (19.4)	70.2 (19.4)	0.8 (1.5)	.61	No	0.04	3,768 (96)
Basic model plus pretest[c]	71.0 (19.4)	70.3 (19.4)	0.7 (1.1)	.55	No	0.03	3,768 (96)
All covariates[d]	71.0 (19.4)	70.4 (19.4)	0.5 (1.1)	.62	No	0.03	3,768 (96)

ATLAST is Assessing Teacher Learning About Science Teaching.

Notes: Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.

a. Benjamini-Hochberg correction used to adjust for multiple comparisons of two outcomes.

b. No covariates except for block dummy indicators, treatment, and site-by-treatment interaction.

c. Basic model plus pretest as an additional covariate.

d. All covariates:

- Student demographic characteristics: sex (male, female), English language learner status (English language learner, fluent English proficient), and race/ethnicity (White, Hispanic, Black, Asian, Other).
- Student pretest measure of outcome variable (ATLAST Test of Force and Motion pretest; standardized grade 7 mathematics scale scores from 2008/09 in lieu of California Standards Test pretest).
- Teacher (random intercept).
- Teacher pretest measure of content knowledge (ATLAST Test of Force and Motion pretest of force and motion).
- Teacher teaching experience, based on five-level scale: beginning (0–2 years), high beginning (3–4 years), middle (5–7 years), high middle (8–10 years), and veteran (11 or more years).
- Teacher undergraduate degree (science, not science).
- Treatment group (intervention, control).
- Site-by-treatment interaction.
- Teacher randomization stratum.
- Missing-value indicators.

Source: Author's analysis of primary data collected for the study.

Table S2. Sensitivity of student impact estimates to different student samples

Measure/sample	Adjusted mean: Intervention Group (standard deviation)	Adjusted mean: Control group (standard deviation)	Difference (standard error)	Unadjusted p-value	Statistical significance after correction[a]	Effect size	Student sample size
ATLAST Test of Force and Motion for students (percent correct)							
Complete cases[b]	52.5 (19.7)	50.8 (19.4)	1.7 (1.0)	.09	No	0.09	4,612 (121)
Pretest and posttest[c]	52.5 (19.8)	50.4 (19.4)	2.1 (1.0)	.04	No	0.11	4,967 (127)
Posttest[d]	52.4 (19.8)	50.3 (19.3)	2.1 (1.0)	.04	No	0.11	5,130 (127)
California Standards Test physical science reporting clusters							
Complete cases[b]	71.5 (19.0)	71.0 (19.1)	0.5 (1.0)	.69	No	0.02	3,273 (96)
Pretest and posttest[c]	71.5 (19.0)	70.9 (19.2)	0.6 (1.1)	.63	No	0.03	3,341 (96)
Posttest[d]	71.0 (19.4)	70.4 (19.4)	0.5 (1.1)	.62	No	0.03	3,768 (96)

ATLAST is Assessing Teacher Learning About Science Teaching.

Notes: Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.

Note: Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.

a. Benjamini-Hochberg correction used to adjust for multiple comparisons of two outcomes.

b. Student sample with valid non-missing pretest and posttest and complete data for all covariates.

c. Student sample with valid non-missing pretest and posttest and missing covariate values replaced with the average of non-missing values.

d. Student sample with valid non-missing posttest and missing pretest and covariate values replaced with the average of non-missing values.

All models were estimated with the full set of all covariates:

- Student demographic characteristics: sex (male, female), English language learner status (English language learner, fluent English proficient), and race/ethnicity (White, Hispanic, Black, Asian, Other).
- Student pretest measure of outcome variable (ATLAST Test of Force and Motion pretest; standardized grade 7 mathematics scale scores from 2008/09 in lieu of California Standards Test pretest).
- Teacher (random intercept).
- Teacher pretest measure of content knowledge (ATLAST Test of Force and Motion pretest of force and motion).
- Teacher teaching experience, based on five-level scale: beginning (0–2 years), high beginning (3–4 years), middle (5–7 years), high middle (8–10 years), and veteran (11 or more years).
- Teacher undergraduate degree (science, not science).
- Treatment group (intervention, control).
- Site-by-treatment interaction.
- Teacher randomization stratum.
- Missing-value indicators.

Source: Author's analysis of primary data collected for the study.

Teacher outcomes

The sensitivity of intervention effects on teacher outcomes was analyzed based on regression models that included varying combinations of covariates (table S3) and different analytic samples (table S4).

Table S3. Sensitivity of teacher impact estimates to different model specifications

	Adjusted mean					
Measure/model	*Intervention Group (standard deviation)*	*Control group (standard deviation)*	*Difference (standard error)*	*Unadjusted p-value*	*Statistical significance after correction*[a]	*Effect size*
ATLAST Test of Force and Motion for Teachers (percent correct)						
Basic model[b]	66.8 (19.2)	57.0 (16.0)	9.8** (3.1)	<.01	Yes	0.61
Basic model plus pretest[c]	65.3 (19.2)	59.2 (16.0)	6.1* (2.2)	<.01	Yes	0.38
All covariates[d]	65.3 (19.2)	59.2 (16.0)	6.2* (2.2)	<.01	Yes	0.38
Confidence in ability to teach force and motion						
Basic model[b]	2.7 (0.3)	2.5 (0.4)	0.2** (0.06)	<.01	Yes	0.46
Basic model plus pretest[c]	2.7 (0.3)	2.5 (0.4)	0.2** (0.04)	<.01	Yes	0.48
All covariates[d]	2.7 (0.3)	2.5 (0.4)	0.2** (0.04)	<.01	Yes	0.49

ATLAST is Assessing Teacher Learning About Science Teaching.
Notes: All models were estimated with teacher sample (n = 133) with valid non-missing posttest data. Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable. *Significantly different from zero at the 0.05 level, two-tailed test. **Significantly different from zero at the 0.01 level, two-tailed test.
a. Benjamini-Hochberg correction used to adjust for multiple comparisons of two outcomes.
b. No covariates except for block dummy indicators, treatment, and site-by-treatment interaction.
c. Basic model plus pretest as an additional covariate.
d. All covariates:

- Teacher sex (male, female).
- Teacher pretest measure of outcome variable (ATLAST Test of Force and Motion pretest of force and motion).
- Teacher baseline confidence in teaching force and motion (teacher survey ratings).
- Teacher teaching experience, based on five-level scale: beginning (0–2 years), high beginning (3–4 years), middle (5–7 years), high middle (8–10 years), and veteran (11 or more years).
- Teacher undergraduate degree (science, not science).
- Treatment group (intervention, control).
- Student academic ability (teacher-aggregated student grade 7 scores on 2008/09 standardized test in mathematics).
- Site-by-treatment group interaction.
- Teacher randomization stratum.
- Missing-value indicators.

Source: Author's analysis of primary data collected for the study.

Table S4. Sensitivity of teacher impact estimates to different teacher samples

Measure/sample	Adjusted mean: Intervention Group (standard deviation)	Adjusted mean: Control group (standard deviation)	Difference (standard error)	Unadjusted p-value	Statistical significance after correction[a]	Effect size	Student sample size
ATLAST Test of Force and Motion for students (percent correct)							
Complete cases[b]	65.3 (19.3)	59.1 (16.1)	6.2* (2.2)	<.01	Yes	0.38	131
Pretest and posttest[c]	65.2 (19.3)	59.0 (16.0)	6.2* (2.2)	<.01	Yes	0.38	132
Posttest[d]	65.3 (19.3)	59.2 (16.1)	6.2* (2.2)	<.01	Yes	0.38	133
Confidence in ability to teach force and motion							
Complete cases[b]	2.7 (0.3)	2.5 (0.4)	0.2** (0.04)	<.01	Yes	0.49	131
Pretest and posttest[c]	2.7 (0.3)	2.5 (0.4)	0.2** (0.04)	<.01	Yes	0.49	132
Posttest[d]	2.7 (0.3)	2.5 (0.4)	0.2** (0.04)	<.01	Yes	0.49	133

ATLAST is Assessing Teacher Learning About Science Teaching.

Note: Data were adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the unadjusted control group standard deviation of the outcome variable.

*Significantly different from zero at the 0.05 level, two-tailed test. **Significantly different from zero at the 0.01 level, two-tailed test.

a. Benjamini-Hochberg correction used to adjust for multiple comparisons of two outcomes.

b. Sample included valid, non-missing pretest and posttest and complete data for all covariates.

c. Sample included valid, non-missing posttest and missing pretest and covariate values replaced with the average of non-missing values.

d. Teacher sample with valid non-missing posttest and missing pretest and covariate values replaced with the average of non-missing values.

All models were estimated with the full set of all covariates:

- Teacher sex (male, female).
- Teacher pretest measure of outcome variable (ATLAST Test of Force and Motion pretest of force and motion).
- Teacher baseline confidence in teaching force and motion (teacher survey ratings).
- Teacher teaching experience, based on five-level scale: beginning (0–2 years), high beginning (3–4 years), middle (5–7 years), high middle (8–10 years), and veteran (11 or more years).
- Teacher undergraduate degree (science, not science).
- Treatment group (intervention, control).
- Student academic ability (teacher-aggregated student grade 7 scores on 2008/09 standardized test in mathematics).
- Site-by-treatment group interaction.
- Teacher randomization stratum.
- Missing-value indicators.

Source: Author's analysis of primary data collected for the study.

Influence of teacher-level covariates

Covariates were varied in three regression models:

- *Basic model*: Included no covariates beyond block dummy indicators, site × treatment interaction, and treatment condition.
- *Basic-plus-pretest model*: Included the variables in the basic model plus baseline test score, and an indicator variable for missing data on the baseline student measure.
- *All covariates*: Included all of the above terms plus the student-level and teacher-level covariates described in chapter 2, and indicator variables for missing data on each applicable covariate.

All models were estimated for the teacher sample with valid, non-missing posttests ($n = 133$). Treatment effects on teachers' content knowledge of force and motion reached statistical significance for all three models. However, the inclusion of the pretest in the impact analysis model (basic model plus pretest) decreased the point estimate from 9.8 to 6.1 and the effect size from 0.61 to 0.38. The differences in estimates when the pretest was included in the basic model likely reflect the significant differences between baseline science scores of intervention and control group teachers (see table 2.5). There were no differences between estimates for the model with pretest only and with all covariates.

With respect to treatment effects on teachers' confidence in their ability to teach force and motion, controlling for covariates did not significantly change the outcome. Treatment effects on confidence reached statistical significance for all three models, with effect sizes of 0.46–0.49.

Influence of analytic teacher sample

Treatment effects were estimated for three models involving different subsets of the teacher data:

- *Complete cases*: Teacher sample with valid, non-missing pretest and posttest and complete data for all covariates ($n = 131$).
- *Pretest and posttest*: Teacher sample with valid, non-missing pretest and posttest and missing covariate values replaced with the average of non-missing values ($n = 132$).
- *Posttest*: Teacher sample with valid, non-missing posttest and missing covariate values replaced with the average of non-missing values ($n = 133$).

All models were estimated with the full set of all covariates. Estimating effects for different analytic samples did not change the outcome with respect to teachers' content knowledge or confidence in their ability to teach force and motion (see table T.4). Treatment effects on teachers' content knowledge of force and motion reached statistical significance for all three models. There were no differences between estimates of impact (6.2), p-values (.05), or effect sizes (.38) for the models with additional missing values. Treatment effects on teacher confidence reached statistical significance for all three models ($p < 0.01$), with effect size of 0.49.

References

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage.

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

Arizona Department of Education. (2010). *2009–2010 State report card.* Phoenix, AZ: Author.

Barnett, J., & Hodson, D. (2001). Pedagogical context knowledge: toward a fuller understanding of what good science teachers know. *Science Education, 85,* 426–453.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a new and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B, 57*(1), 1289–1300.

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

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: CCSSO.

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

California Department of Education. (2011a). 2009 STAR test results. Retrieved January 7, 2011, from http://star.cde.ca.gov/star2009/SearchPanel.asp?ps=true&lstTestYear=2009.

California Department of Education. (2011b). California Standards Tests Technical Report: Spring 2010 Administration. Retrieved July 22, d011, from http://www.cde.ca.gov/ta/tg/sr/technicalrpts.asp.

Carlsen, W. S. (1991). Subject-matter knowledge and science teaching: a pragmatic perspective. In J. Brophy (Ed.), *Advances in research on teaching* (Vol. 2, pp. 115–124). Greenwich, CT: JAI 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, CT: 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.

Daehler, K. R., Shinohara, M., & Folsom, J. (2011). *Making Sense of SCIENCE$^{TM:}$ Force and motion for teachers of grades 6–8.* San Francisco, CA: WestEd.

Delta Education (2010). *FOSS: Full option science system (3rd edition)*. 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.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*(2), 81–112.

Donner, A. N., & Klar, N. (2000). Design and analysis of cluster randomization trials in health research. London: Arnold.

Driver, R., Guesne, E., & Tiberghien, A. (Eds.). (1985). *Children's ideas in science*. Milton Keynes, UK: Open University Press.

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.

Dynarski, M., Moore, M., Rosenberg, L., James-Burdumy, S., Deke, J., & Mansfield, W. (2004). When schools stay open late: the national evaluation of the 21st Century Community Learning Centers Program, new findings. Final report. Princeton, NJ: Mathematica Policy Research, Inc.

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

Fox, J. (2002). *Bootstrapping regression models*. *Appendix to an R and S-PLUS companion to applied regression*. Retrieved November 3, 2010 from http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf

Franke, M. L., Carpenter, T., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal, 38*, 653–690.

Fulp, S. (2002). 2000 national survey of science and mathematics education: status of middle school science teaching. Chapel Hill, NC: Horizon Research, Inc.

Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. (2001). What makes professional development effective? Analysis of a national sample of teachers. *American Educational Research Journal, 38,* 915–45.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Oxford University Press.

Hapkiewicz, A. (1999). Naïve ideas in earth science. *Michigan Science Teachers Association Journal, 44*(2), 26–30. Retrieved December 1, 2010, from http://www.msta-mich.org

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

Hawker Brownlow. (2010). *STEM-CIP: Science/technology/engineering/mathematics curriculum integration program.* Moorabbin, Victoria, Australia: Author. Available at http://www.currtechintegrations.com/stem-cip.php.

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. (in press, to appear March 2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*.

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 WestEd and Stuart Foundation.

Heller, J. I., Shinohara, M., Miratrix, L., Rabe-Hesketh, S., & Daehler, K.R. (2010, March). *Learning science for teaching: effects of professional development on elementary teachers, classrooms, and students.* Paper presented at the 2010 Conference of the Society for Research on Educational Effectiveness, Washington, DC.

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. C., & Ball, D. L. (2004). Learning mathematics for teaching: results from California's mathematics professional development institutes. *Journal for Research in Mathematics Education. 35*(5), 330–351.

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.

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

Knapp, M. S., McCaffrey, T., & Swanson, J. (2003, April). *District support for professional learning: what research says and has yet to establish.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research, 75*(4), 491–530.

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. (2006). *NEA research best practices: professional community and professional development in the learning-centered school.* Washington, DC: National Education Association. Retrieved October 23, 2010, from http://www.nea.org /tools/30380.htm

Murray, D. M. (1998). *Design and analysis of group randomized trials.* New York: Oxford University Press.

National Research Council. (1996). *National science education standards.* Washington, DC: National Academies Press. Retrieved October 23, 2010, from http://www.nap.edu/catalog.php?record_id=4962

National Staff Development Council. (2001). *National Staff Development Council's standards for staff development (revised)*. Retrieved October 23, 2010, from http://www.learningforward.org/standards/index.cfm

Partnership for 21st Century Skills. (2008). *21st century skills, education and competitiveness: a resource and policy guide.* Tucson, AZ: Author. Retrieved October 23, 2010, from http://www.p21.org/index

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials.* Washington, DC: U.S. Department of Education, National Center for Educational Evaluation (NCEE 2009-0049). Retrieved December 1, 2010, from http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20090049

The R Foundation for Statistical Computing. (2009). *R version 2.9.2 (2009-08-24)*. Vienna, Austria: Department of Statistics and Mathematics. Retrieved December 1, 2010, from http://cran.r-project.org/

Raudenbush, S. W. (1997). Statistical analysis and optimal design in cluster randomized trials. *Psychological Methods, 2,* 173–185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Sawchuk, S. (2010). Professional development for teachers at crossroads. *Education Week, 30*(11), s2–s4. Retrieved November 10, 2010, from http://www.edweek.org

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.

Schochet, P. Z. (2005). Statistical power for random assignment evaluations of education programs. Princeton, NJ: Mathematica Policy Research, Inc.

Schochet, P. Z. (2008). Guidelines for multiple testing in experimental evaluations of educational interventions. Princeton, NJ: Mathematica Policy Research, Inc.

Schweingruber, H. A., & Nease, A. A. (2000, April). *Teachers' reasons for participating in professional development programs: do they impact program outcomes?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Shinohara, M., Daehler, K. R., & Heller, J. I. (2004, April). *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, BC, Canada.

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.

Smith, S. P., & Banilower, E. R. (2006a, April). *Measuring middle grades students' understanding of force and motion concepts: insights into the structure of student ideas.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco.

Smith, S. P., & Banilower, E.R. (2006b, April). *Measuring teachers' knowledge for teaching force and motion concepts.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco.

Terrell, N. (2007). STEM occupations: high-tech jobs for a high-tech economy. *Occupational Outlook Quarterly, Spring,* 26–33. Washington, DC: Office of Occupational Statistics and Employment Projections, U.S. Department of Labor.

Tharp, R. G., Estrada, P., Dalton, S. S., & Yamauchi, L. (2000). *Teaching transformed: Achieving excellence, fairness, inclusion, and harmony*. Boulder: Westview Press.

Torres, H. N., & Zeidler, D. L. (2002). The effects of English language proficiency and scientific reasoning skills on the acquisition of science content knowledge by Hispanic English language learners and native English language speaking students. *Electronic Journal of Science Education, 6*(3), Article Four.

U.S. Department of Education. (2004). *Trends in international mathematics and science study: science items.* Washington, DC: U.S. Department of Education, National Center for Educational Evaluation. Retrieved August 13, 2010, from http://nces.ed.gov/timss/educators.asp

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.

von Hippel, P. T. (2007). Regression with missing Ys: an improved strategy for analyzing multiply-imputed data. *Sociological Methodology, 37,* 83–117.

Wallace, M. R. (2009). Making sense of the links: professional development, teacher practices, and student achievement. *Teachers College Record*, *111*(2), 573–596. Retrieved November 9, 2010, from http://www.tcrecord.org

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.

Weiss, I. R., Banilower, E. R., McMahon, K. C., & Smith, P. S. (2001). *Report of the 2000 National Survey of Science and Mathematics Education.* Chapel Hill, NC: Horizon Research, Inc.

Weiss, I. R., Gellatly, G. B., Montgomery, D. L., Ridgway, C. J., Templeton, C. D., & Whittington, D. (1999). *Executive summary of the local systemic change through teacher enhancement year four cross-site report.* Chapel Hill, NC: Horizon Research, Inc.

What Works Clearinghouse. (2008). *Procedures and standards handbook (version 2.0).* Retrieved October 23, 2010, from http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf

White, I. R., & Thompson, S. G. (2005). Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine, 24*(7), 993–1007.

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: an examination of research on contemporary professional development. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education, 23,* 217–234. Washington, DC: American Educational Research Association.

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


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA