Educational Evaluation and Policy Analysis
September 2017, Vol. 39, No. 3, pp. 511–536
DOI: 10.3102/0162373717697842

Inside the Black Box: Examining Mediators and Moderators of a Middle School Science Intervention

Laura M. Desimone
University of Pennsylvania
Kirsten Lee Hill
Orleans Parish School Board

We use data from a randomized controlled trial of a middle school science intervention to explore the causal mechanisms by which the intervention produced previously documented gains in student achievement. Our study finds that implementation fidelity, operationalized as a measure of the frequency of implementation of the cognitive science principles taught in intervention teachers' professional development, helped to explain student achievement effects. Integrating findings from a structural equation model using data from 10,281 students and 124 teachers with a small subsample of teacher interviews, we also found that the intervention may work partly through fostering better classroom management and collaborative discussions that elevated practice. Furthermore, our results have implications for informing decisions about how to balance a focus on increasing teacher content knowledge, on one hand, and providing explicit pedagogical strategies linked to the curriculum, on the other. We additionally found that lower achieving classrooms had lower implementation scores, likely due to factors that hindered teachers' ability to use the cognitive science principles, such as less science class time or needing to adapt content to meet the needs of struggling students. Our study highlights the importance of anticipating—and calibrating interventions to—the contextual complexities of real-life classrooms, and it identifies several factors with the potential to contribute to improved design and evaluation of such interventions.

Keywords: *educational reform, instructional practices, middle schools, mixed-methods, professional development, structural equation modeling, student cognition, survey research*

Explaining the Effects of School and Classroom Interventions

EDUCATIONAL policy researchers have increasingly recognized that in studying the effectiveness of an intervention, it is important to understand not only *if* the intervention increased student achievement but also *why* it did or did not work (O'Donnell, 2008). By exploring the causal mechanisms by which an intervention succeeded or failed, researchers can extend beyond evaluations of efficacy to make theoretical contributions and practical improvements (Ruiz-Primo, 2005). Furthermore, because educational interventions are implemented in complex environments, it is important to examine how sensitive effects are to both context and heterogeneous implementation to guide theory and identify the conditions under which the intervention was more or less effective (e.g., Cook, 2002; Duncan & Magnuson, 2003; Heckman & Smith, 1995; Raudenbush, 2005).

In this study, we analyze implementation in the context of a clustered randomized control trial (RCT) of a middle school science intervention that administered professional development (PD) emphasizing principles of cognitive science (CS). Previous research on this intervention demonstrated that the CS intervention was successful in

achieving small gains in student achievement compared with a business-as-usual control group (Scull, Porter, Merlino, & Massey, 2017; Yang, Porter, Merlino, & Massey, 2017). Contributing to our understanding of how implementation and contextual factors may influence the success of school/classroom-based interventions, we provide insights about the mechanisms by which the intervention worked and the contexts in which it was more successful.

Research Questions: The Role of Implementation

After establishing that implementation does matter for effects on students, we explore mechanisms that may explain why some teachers had higher implementation scores than others. Using a combination of survey, achievement, and interview data, we look inside the "black box" of implementation to evaluate several mechanisms by which the intervention may have yielded gains in student achievement.

First, we ask, *to what extent does classroom implementation of the intervention mediate effects on student achievement?* Here, we create a measure of implementation based on the frequency with which teachers used the intervention's CS principles (taught in PD) in their teaching of science in the intervention classrooms. Second, because the intervention also relayed science content taught in the middle school curriculum, we ask, *to what extent does teacher content knowledge mediate the effects of the intervention on student achievement?* Previous research has underscored the important roles of teacher and classroom context in implementation (e.g., Holme & Rangel, 2012; McLaughlin, 2005); thus, in this study, we also ask, *how and in what ways is teachers' implementation of the intervention influenced by teacher experience, subject expertise (as measured by college major), and prior classroom achievement?* Many interventions aim to raise achievement in low-achieving classrooms, so understanding how contextual factors may hinder the implementation of interventions has implications for improving how we support schools and teachers in their efforts to engage in improved practices. We use data from interviews with participating teachers to help understand and explain our findings from the quantitative analyses, further exploring why the intervention may have worked, and why it may be more difficult to implement in classrooms with lower prior achievement.

The Research Base for the Intervention

Prior research has suggested that using CS principles in teaching will increase student achievement in middle school science (Bransford & Schwartz, 1999; Chi, 2005; Hegarty, Kriz, & Cate, 2003). Teachers in our intervention study received PD on how to implement three CS principles—contrasting cases, visualization, and spaced testing. Teachers additionally received a detailed *Cognitive Science Casebook* that provided written materials and classroom activities based on these principles.

Instruction that uses contrasting cases asks students to read about, analyze, and compare two related ideas, events, or phenomena—for example, evaporation from a puddle of rainwater and steam rising from a boiling kettle. This type of instruction helps students to identify and understand key concepts, make inferences, and explain their reasoning (Chi, 2005; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, De Leeuw, Chiu, & LaVancher, 1994). Use of contrasting cases has been shown to help students learn new materials and concepts (Bransford & Schwartz, 1999; Gentner, Loewenstein, & Thompson, 2003; Kurtz, Miao, & Gentner, 2001; Schwartz & Bransford, 1998). The second principle, visualization, targets the difficulties middle school students have analyzing and interpreting visual representations such as charts, diagrams, and graphs (Hegarty et al., 2003). Instruction focused on visualization strategies provides students with direct instruction and scaffolding to guide them through various depictions, such as a line diagram depicting a water cycle. The third principle, spaced testing, is based on research that has shown that regularly testing students on material is more effective for students' knowledge retention than simply reviewing the material (Roediger & Karpicke, 2006). To these ends, the intervention provided teachers with quizzes that they were to administer to their students on particular days. These quizzes covered content that was recently taught as well as content from earlier in the unit.

Classroom Implementation

The PD intervention tested in the RCT was designed to increase student achievement through the application of the principles of CS described in the previous section. As such, it is expected that the success of the intervention is related to the fidelity of teachers' implementation of these principles (Cordray & Pion, 2006; Dane & Schneider, 1998; Moncher & Prinz, 1991).

Research on implementation fidelity and how it relates to the effectiveness of interventions is lacking, particularly in the field of education (O'Donnell, 2008). Despite an increased emphasis on the importance of measuring implementation in effectiveness studies (e.g., National Research Council [NRC], 2004; Winters, Wise, & Towne, 2004), implementation is still not often reported in effectiveness studies (e.g., U.S. Department of Education [USDOE], 2015), and in cases where it is, it tends to be a standalone measure that is not considered in interpreting effectiveness outcomes (Lendrum & Humphrey, 2012; NRC, 2004). In a review of journals that published high-impact findings (i.e., that appear in articles that are frequently read and cited) from general and special education interventions between 2005 and 2009, Swanson, Wanzek, Haring, Ciullo, and McCulley (2011) found that out of 76 articles reviewed, only 67% provided fidelity procedures and only 47% provided quantitative data on implementation fidelity.

The process of examining implementation fidelity is complicated by several factors. First, there is not widespread agreement on how to define or measure implementation fidelity, which results in substantial inconsistencies in measurement approaches across studies (O'Donnell, 2008). Furthermore, local adaptation—which by its very nature causes variation in implementation fidelity—could be either beneficial or detrimental to an intervention (Berman & McLaughlin, 1976; Blakely et al., 1987; Durlak, 2010; Dusenbury, Brannigan, Falco, & Hansen, 2003). The tension between implementation fidelity and local adaptation has deterred the What Works Clearinghouse (WWC) from developing standards for measuring implementation fidelity. The WWC of the U.S. Department of Education (USDOE, 2015) explains that one might expect that competing demands placed on teachers result in variation in implementation and notes that this variation ought to be observed and described. For instance, teachers may adapt an intervention due to a lack of resources or to meet the needs of particular students (McHugo et al., 2007; O'Donnell, 2008; USDOE, 2015). For these reasons, the WWC decided not to pursue standards for measuring or evaluating implementation fidelity. Instead, it highlighted the importance of understanding whether and why or why not an intervention is effective in a real-world setting (USDOE, 2015).

In our study, we examine implementation fidelity and adaptation in an integrative way (Webster-Stratton, Reinke, Herman, & Newcomer, 2011). Rather than viewing implementation fidelity as undercutting the potential value of local adaptation, we see it as providing a more systematic way to contextualize the results of effectiveness studies. By using implementation fidelity as a tool for contextualizing results, we allow for a more comprehensive understanding of how and why an intervention did or did not work, thereby leveraging the strengths of measuring implementation fidelity to explain adaptation (Desimone, McMaken, & Cherng, 2010; Summerfelt, 2003).

There are many dimensions of implementation fidelity, and an array of terminology is used to describe these dimensions (Hulleman & Cordray, 2009). For our study, we draw on Dane and Schneider's (1998) five criteria for implementation fidelity, conceptualizing implementation fidelity as the degree to which the activities, materials, and procedures that comprise the intervention are administered as they were intended. Although Dane and Schneider (1998) recommend measuring all five components of implementation fidelity (i.e., adherence, duration, quality of delivery, participant responsiveness, and program differentiation), it is unclear whether fidelity on all five dimensions is necessary for a program to be effective in meeting its goals, whether student achievement, decreased drug use, or so forth (Dusenbury et al., 2003). In this study, we focus on one dimension of fidelity—teachers' *adherence* to the intervention. We measure adherence quantitatively with constructs from a teacher survey, and consider this measure of implementation fidelity as a potential mediator of the intervention's effectiveness. For a more in-depth discussion of implementation

fidelity and the development of a quality measure for adherence as used in this study, see Desimone, Richards, and Hwang (2013).

Although many curriculum interventions are designed to improve the learning outcomes of low-achieving students, they are implemented in a variety of classrooms, and there is little research that compares implementation in higher and lower achieving classrooms to better understand potential challenges to implementation (Berends, Chun, Ikemoto, Stockly, & Briggs, 2002). With the goal of providing additional insights into the often-elusive contextual factors that may influence implementation, in our analysis, we consider the extent to which factors including teacher experience, teacher major, and class prior achievement affect implementation fidelity (Durlak & DuPre, 2008).

The Role of Teacher Experience and Content Knowledge

While the intervention centered on implementation of CS principles, teacher expertise—both through content knowledge and experience—is another important pathway that may improve student achievement, making it a crucial variable to account for in studies that seek to explain teacher effects on student achievement (Hill, Rowan, & Ball, 2005; Schulman, 1986).

Even though teacher content knowledge has not received as much attention in the sciences as it has in mathematics (see Diamond, Maerten-Rivera, Rohrer, & Lee, 2014), researchers have established a link between teachers' knowledge of science and their students' science achievement (e.g., Fleer, 2009; Shallcross, Spink, Stephenson, & Warwick, 2002; Supovitz & Turner, 2000). As in mathematics, there seems to be a consensus that elementary and middle school teachers, on average, do not have the strong science knowledge needed to foster 21st-century science learning for students (e.g., Jüttner, Boone, Park, & Neuhaus, 2013; Nowicki, Sullivan-Watts, Shim, Young, & Pockalny, 2013). When teachers are well versed in science, however, research finds they are better able to support deep and meaningful student learning across multiple science domains (Diamond et al., 2014).

Research on the role of teacher experience in student learning is mixed. Teachers' effectiveness at increasing student achievement improves during the first few years on the job (Clotfelter, Ladd, & Vigdor, 2007; Harris & Sass, 2007; Loeb, Beteille, & Kalogrides, 2012); however, teachers vary considerably in what they learn over time and how they translate that knowledge into practice, and there is much to be learned about how teacher knowledge and practice translates to student learning, such as how long it takes, what types of knowledge translate to which practices, and so on (e.g., Domitrovich et al., 2009; Garet et al., 2010; Gowlett et al., 2015; Penuel, Fishman, Yamaguchi, & Gallagher, 2007).

There is a small amount of research on the role of teacher content knowledge and experience in predicting the frequency and quality of intervention implementation. For example, researchers have found that novice teachers are more likely to support reforms, such as instructional interventions, than are veteran teachers (e.g., Berends, 2000), and that interventions that ask teachers to perform complex conceptual tasks are more likely to be well executed by teachers with higher content knowledge in the subject area (e.g., Hill et al., 2005; Metzler & Woessmann, 2012; Sadler, Sonnert, Coyle, Cook-Smith, & Miller, 2013). In our analysis, we seek to better understand the roles of teacher content knowledge and experience in explaining variation in student achievement attributed to the CS intervention.

Research Design

The data for our implementation analysis are drawn from a broader three-armed RCT that occurred in phases from 2009 to 2012. The RCT tested the effectiveness of a CS intervention against a business-as-usual control and a science content knowledge intervention, on two cohorts of three different science content units in seventh and eighth grade (see Figure 1). The intervention we report on here was conducted in a large urban city in the northeastern United States. To promote teacher collaboration and minimize contamination across conditions, a cluster randomized trial was conducted—schools were randomized into one of the three arms: the CS arm, the content arm, and the control arm.

Although both the CS-based and content-based treatment groups were hypothesized to improve

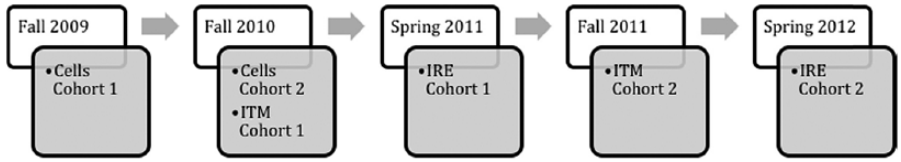

FIGURE 1. *Unit implementation timeline.*
Note. The three units were Cells, Heredity, and Classification (Cells); Inside the Restless Earth (IRE); and Introduction to Matter (ITM).

achievement over the business-as-usual control group, the CS-based group represents the focal point of the larger RCT; the content-based group was included as a benchmark for comparing a more traditional content-based PD approach with the CS-based approach. That is, because the CS intervention's adaptations to the curriculum (inclusion of principles of CS) cannot be taught in isolation (i.e., without also focusing on science content), in an effort to parse out what changes in student achievement came about through increased content knowledge as opposed to exposure to the principles of CS, we also included a treatment condition that provided science content PD without the CS principles. We expected teachers in the content arm to increase their content knowledge more than control or CS teachers. Although we did not expect significant increases in CS teachers' content knowledge, we wanted to be able to compare their content knowledge with the content arms', to enable us to eliminate content knowledge as the mechanism by which effects did or did not occur.

Teachers in both treatment conditions participated in summer PD specifically targeting three units of *Holt Science and Technology,* a widely adopted, traditional textbook-centered curriculum. The three units were Cells, Heredity, and Classification (hereafter Cells); Inside the Restless Earth (IRE); and Introduction to Matter (ITM). We chose to study three units to establish generalizability of this approach to science instruction across various content areas. The three units chosen for the intervention are standalone units that are congruent with each other (thus assessable on a common measure) and were aligned to the state's science standards.

Teachers in the CS treatment condition received PD focusing on three core CS principles that have previously been linked to increases in student achievement: contrasting cases, visualization, and spaced testing (Alfieri, Nokes-Malach, & Schunn, 2013; Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Cromley et al., 2010; Newcombe, 2013). Teachers in the content treatment condition received PD in the general topic areas covered in middle school science, and teachers in the business-as-usual control group participated in any normal PD offered by their school and/or district. The summer sessions for the CS condition focused on specific modifications to the science curriculum as well as how and why these principles of CS should be integrated into the curriculum during classroom lessons. In addition to the PD, teachers in the CS condition received the researcher-constructed *Cognitive Science Casebook* of approximately 500 pages with curriculum modifications. The Casebook provided step-by-step lesson plans, including detailed explanations, warm-ups, activities, and PowerPoint slides, for the teacher to follow when teaching the target units. These detailed lessons and activities reflected the CS principles and were aligned to the Holt curriculum. The *Casebook*, which was to be used in place of the districts' planning guides, was aligned to district curricular guidelines.

PD for teachers in the Content condition focused on science content, in particular topics that the state highlighted as eligible for statewide testing. Teachers in the Content condition received a binder of materials akin to the *Casebook*; however, their materials were strictly content related (Massey, Cleland, & Mandel, 2013).

For both treatment conditions, PD was taught by science museum professionals, university professors, science researchers, and high school content area teachers *who specialized in the given content unit*. Follow-up sessions during the academic year were modeled as professional learning communities (PLCs) for both treatment conditions—CS and Content. The PLCs gave teachers an opportunity to share their successes and difficulties with instruction and offer

guidance and support to one another. For the CS condition, curricular modifications continued to be presented as part of the PLCs.

All PD provided as part of the intervention reflected five key features of high-quality PD that have been shown in rigorous empirical studies to be related to changes in instruction: PD was focused on content, included active learning opportunities for teachers, was coherently integrated into the curriculum, provided a substantial number of sustained contact hours, and included collective participation of teachers from the same subject (e.g., Desimone & Garet, 2015; Garet, Porter, Desimone, Birman, & Yoon, 2001; Penuel et al., 2007). Specifically, teachers in the CS and Content treatment conditions participated in 2.5 days of summer PD *for each unit* prior to the unit beginning (see Figure 1 for our implementation timeline). Teachers also participated in four follow-up PLCs during the semester in which they implemented a given unit; these lasted 2 hours each. In total, *for each unit* they participated in, in the first year of the study, teachers could receive up to 18 hours of summer PD and 8 hours of PLCs (2 hours per PLC meeting; PLC meetings took place on a monthly basis for 4 months). Cohort 2 teachers who were returning to teach the same unit did not have to repeat the summer PD but did continue to participate in the monthly PLC meetings during their second year of implementation. The time spent in PD was held constant across the CS and Content groups. Previous analyses of this intervention found that hours of PD participation are unrelated to implementation fidelity (Scull et al., 2015).

Data Collected

In the spring of 2011 and 2012, we administered surveys to teachers, which asked about their classroom science instruction. Surveys were Web based, but we made hard copies available to several teachers who preferred to complete them via paper and pencil.

Across all three treatment arms, we targeted 508 teachers; 339 completed the surveys, for a response rate of 66.7%. For the CS treatment condition, we targeted 181 teachers; 127 completed the surveys, for a response rate of 70.2%. For the Control condition, we targeted 184 teachers; 124 completed the surveys, for a response rate of 67.4%. For the Content condition, we targeted 153 teachers, and 88 completed the surveys, for a response rate of 61.5%. Although randomization usually results in baseline equivalence, this is not guaranteed, and so we compared teachers in each arm of the study on key teacher characteristics. We found that teachers in the CS treatment condition had significantly less experience and were significantly less likely to be White, compared with the other two arms. There were no significant differences in response rates by respondent group.

We followed up with teachers who attrited from the study and found that the top two causes for attrition (which accounted for almost 50% of attrition) were that teachers were "assigned to teach another grade" or "assigned to teach another subject." We also examined the relationship between teachers' propensity of attrition and student achievement and did not find the relationship to differ by arm. Given these findings, we concluded that attrition was not likely to be biased in ways directly related to implementation and student achievement (Yang, Porter, Merlino, & Massey, 2017).

Teachers were eligible to participate in both cohorts of each of the three science units. Teachers were also able to teach multiple units simultaneously. Teachers completed PD, content knowledge assessments, and implementation surveys for *each cohort and each unit* that they taught. This means that if, for instance, a teacher taught Cells for Cohort 1 and Cohort 2, IRE for Cohort 2, and ITM for Cohort 2, then he or she could have three content knowledge assessment scores and up to four completed implementation surveys.

We conducted teacher interviews with a subsample of 14 teachers implementing the IRE unit. Teachers were selected from each of the three conditions; six of the interviews were with teachers in the CS treatment arm. The CS arm was overrepresented in the sample because the focus of the study is the use of CS principles. The criteria we used to choose teachers were exposure to the intervention and extent of participation in the PD. Specifically, we targeted teachers who were in the second year of participating in the study, teachers who had the highest number of PD hours for the unit, and teachers who taught multiple sections of science, because those teachers had more opportunities for implementation. Research has shown that exposure to an intervention is related

TABLE 1

Teacher Interview Characteristics

Teacher	Years of experience	Undergraduate major	Average teacher content knowledge score (IRE[a])[b]	Average implementation score (IRE[c])[d]
Chloe	3	Geology	16	2.90
Betty	8	Education	19	2.62
Gina	28	Education	19	2.84
Lisa	2	Biology	20	3.21
Ryan	2	Social science	22	2.78
Jack	3	Geology	22	3.13

[a]Average score for teachers on the content knowledge test for the Inside the Restless Earth (IRE) unit.
[b]Teacher content knowledge scores for full sample: $M = 19.12$; $SD = 3.08$; Range = 8–25.
[c]Average score for teachers on the implementation fidelity survey for the Inside the Restless Earth (IRE) unit
[d]Cognitive science principles implementation score for IRE for full sample: $M = 2.69$; $SD = 0.30$; Range = 2.13–3.56.

to effects of the intervention on knowledge and instruction (Yoon, Duncan, Wen-Yu Lee, Scarloss, & Shapley, 2007). In addition, research finds that teachers are still learning how to implement during the first year of an intervention (Borman, Gamoran, & Bowdon, 2008; Fullan, 1991). See Table 1 for a list of teachers interviewed from the CS arm, and their characteristics. We use pseudonyms to protect teachers' confidentiality.

The interviews lasted 45 minutes on average. For each interview, we used a detailed interview protocol that asked questions about teachers' experiences in the intervention PD, a description of how the teacher taught a particular lesson, and factors that facilitated or interfered with their implementation of the intervention.

Measures

Implementation

Survey data (pooled from each of the cohorts) were used to construct an overall measure of implementation fidelity *for each teacher*, for each unit he or she taught in each cohort. As discussed in the preceding paragraph on data collected, teachers were eligible to participate in both cohorts of each of the three science units, *and* teach multiple units simultaneously. This means that if, for instance, a teacher taught Cells for Cohort 1 and Cohort 2, IRE for Cohort 2, and ITM for Cohort 2, then he or she could have up to four completed implementation surveys, and therefore, four separate measures of implementation fidelity. This measure of implementation fidelity is comprised of three subscores aligned to the intervention's three principles of CS—contrasting cases, visualization, and spaced testing. Our unit-specific surveys asked detailed questions about each of the CS components of the intervention for specific topics that made up the unit. Language in the surveys was generic (i.e., not specific to the CS intervention arm) to ensure that teachers across all arms could answer the questions. To review the questions, items, and answer categories for each construct, see the appendix. The Cells unit survey is used for demonstration purposes but questions were consistent across units.

We developed scales of 8 to 11 items for each of the three implementation subconstructs: compare and contrast (Cronbach's $\alpha = .763$), spaced testing (Cronbach's $\alpha = .695$), and visualization (Cronbach's $\alpha = .396$). The visualization composite had a low alpha, not unexpected with so few items, but it has strong face validity—looking at the appendix, one can see that the survey items for the visualization composite center around analysis and interpretation of visual representations such as charts, diagrams, and graphs—the definition of our visualization construct (Hegarty et al., 2003). Then, using these items, we averaged the responses across variables (responses coded 1–4 or 1–3, depending on the scale) to create composite implementation scores that measure the frequency with which teachers implemented each of the three principles of CS. Higher scores indicate higher implementation

fidelity. Diagnostic testing showed that the three components operated in similar ways, so we averaged the implementation subscores to create an overall measure of implementation fidelity.

Student Achievement

To measure student achievement, teachers administered researcher-constructed end-of-unit tests. To create these tests, researchers drew test items from a pool of publicly available items from state tests, the National Assessment of Educational Progress, and the Trends in International Mathematics and Science Study. To ensure that these tests were aligned to the content of each unit, analysts used the Surveys of Enacted Curriculum (SEC) for science. The science SEC is a matrix of science-specific content topics by student expectations (cognitive demands). The five dimensions of student expectations/cognitive demands measured by the SEC are memorize, perform procedures, demonstrate understanding, conjecture/analyze, and solve nonroutine problems (Porter, 2002; see also Council of Chief State School Officers, 2005; Polikoff & Porter, 2014; Polikoff, Porter, & Smithson, 2011). To create the achievement tests, instructional materials are first analyzed using the SEC (e.g., curriculum is scored using the SEC matrix of content topics by cognitive demands). Such an analysis lays out detailed specifications that describe the domain of content to be covered by the tests. Items for the test are then chosen so that they are a representative sample of the domain and, when grouped together, are maximally aligned to unit-specific content (Polikoff et al., 2011; Porter, Polikoff, Barghaus, & Yang, 2013). The researcher-constructed tests were demonstrated by prior research to be both valid and reliable (e.g., internal reliabilities range from .65 to .69; see Scull et al., 2015; Yang et al., 2017, for details about the validity and reliability of these tests). In terms of missing student data, there was 16% missing for the CS group and 15% missing for the Control. Missing data procedures will be discussed in the forthcoming section on analytic approach.

Administrative Data

In our model, we included student-level demographics and prior-achievement data. Specifically, we included the following binary variables: race (Black, Hispanic, Asian, and Other—with White as a reference category), sex, economic disadvantage, and English language learner status. We also included a continuous variable for prior achievement as measured by statewide standardized tests. These data were provided to us by the participating districts.

Teacher Experience and Knowledge

Teachers reported years of teaching experience on our implementation survey. We consider teacher experience an important control variable in any study of classroom interventions effects on implementation or student learning. For our analyses, we created a binary "novice teacher" variable that defined a novice teacher as anyone with 2 or fewer years of teaching experience. We made this decision based on both theoretical and empirical grounds. There is a considerable literature on substantive differences between new and veteran teachers in their responses to PD, and their implementation of new reforms, we believe this distinction is important, given our research questions focused on how teachers respond to an intervention. Previous research has shown that the trajectory of new teacher learning occurs most dramatically after 2 to 3 years of teaching (Clotfelter et al., 2007; Harris & Sass, 2007; Loeb et al., 2012), and also that teachers in their first years of teaching have been shown to differ significantly in terms of their instruction, classroom management, and reaction to PD (Feiman-Nemser, 2012; Luft et al., 2015). As in other studies, our initial exploration of a continuous teacher experience control variable yielded null results, as we might expect, because a continuous variable assumes a strict linear relationship, which contradicts previous research, which shows substantial gains after the first 2 to 3 years. Thus, we concluded that creating a "novice" variable that reflected whether the teacher was in his or her first 2 years of teaching would make an appropriate control variable for this study.

Our implementation surveys also collected data on teachers' educational background. From these data, we created a binary "STEM major" variable. All science majors (general, physics, chemistry, biology, and so on), all math majors, and any engineering majors were identified as

STEM.[1] In addition to these two somewhat distal proxies for content knowledge, we developed assessments of teacher content knowledge *for each curricular unit* to provide a more closely aligned measure of content knowledge (Becker & Aloe, 2008). That is, the content knowledge test reflected content taught in the three units in our study—Cells, IRE, and ITM. Thus, the domains on the test were biological science, physical science, and geology. These tests were administered at the conclusion of the summer PD sessions.

To create this measure of teacher content knowledge, items were taken from the PRAXIS, Diagnostic Science Assessment for Middle School Teachers, Geo Science Concept Inventory, and the MOSART test. The items were content analyzed again using procedures from the SEC. The length of the test was kept short to minimize the burden to teachers. To align the test to the content taught in the PD, an algorithm was used that results in maximal alignment (Porter et al., 2013). The internal reliability of our content knowledge test ranged from .62 to .64 (Scull et al., 2015; Yang et al., 2017). The response rate for the unit-specific, content knowledge test was approximately 97% for the CS treatment condition and approximately 94% for the Control condition. Teacher quality has been operationalized in multiple ways, including college major, course taking, years of experience, certification, and test scores (Sadler et al., 2013). Although evidence is mixed in relating these variables to instructional quality and student learning (e.g., Wilson, Floden, & Ferrini-Mundy, 2002), our choice to use college major, experience (e.g., not novice), and a knowledge assessment is guided by research documenting their relationship with improved instruction and student learning (Clotfelter et al., 2007; Goldhaber & Brewer, 1997, 2000; Hill et al., 2005; Rowan, Chiang, & Miller, 1997).

Analytic Approach

We examine data on the achievement of 10,281 students enrolled in the classes of 124 teachers in the study (control n = 60, treatment n = 64). Characteristics of the sample of teachers in the control and treatment conditions are provided in Table 2. Characteristics of the analysis sample of teachers and students did not differ significantly across treatment arms in any discernible pattern (Yang et al., 2017). Prior research on this study has validated the randomization and found that overall and differential attrition biases were within acceptable levels as specified by the WWC (Porter et al., 2013).

We address the study's research questions through estimating a series of structural equation models using the Mplus software. Structural equation modeling allows us to assess the indirect mediating effects of implementation and content knowledge on the effectiveness of the intervention as measured by student achievement (Fairchild, MacKinnon, Taborga, & Taylor, 2009; Kline, 2011; T. D. Little, Card, Bovaird, Preacher, & Crandall, 2007; Muthén & Muthén, 1998–2012). In addition, we examine how teachers' implementation fidelity is related to the characteristics of the classrooms they teach. Our conceptual framework is depicted in Figure 2.

The dependent variable in our model is student achievement as measured by our researcher-constructed aligned science test, administered to students at the end of each unit. Exogenous variables include treatment condition, teacher characteristics, student characteristics, and prior-class achievement. Endogenous variables include measures of implementation (overall composite of contrasting cases, spaced testing, and visualization) and teacher content knowledge. Exogenous variables are assumed to be correlated, and the model is estimated conditioned on exogenous variables (see Kline, 2011; Muthén & Muthén, 1998–2012). By default, Mplus treats missing data using full information maximum likelihood. Maximum likelihood is recognized as a strong approach to handling missing data as prior work demonstrates it produces unbiased parameter estimates whether data is Missing at Random (MAR) or Missing Completely at Random (MCAR) (Allison, 2001; Enders & Bandalos, 2001; see Figure 3 for our model specification).

Teachers were eligible to administer student assessments and take content knowledge tests and implementation surveys for up to three different academic units over the span of 3 years (two cohorts per unit). To account for this clustering in our data, we use a complex estimation strategy in structural equation modeling that accounts for clustering at the unit by classroom level (Muthén

TABLE 2

Characteristics of Teachers in Control, Content, and CS Treatment Conditions

	Control		Content		Cognitive science	
	n	%	*n*	%	*n*	%
Gender						
Female	47	35	39	33	53	33
Male	87	65	81	67	105	67
Race						
White	108	61	86	58	130	72
Black	48	24	59	36	26	13
Asian	0	0	0	0	1	2
Other	11	2	2	1	7	4
Multiracial	8	11	0	3	9	9
Ethnicity						
Hispanic	12	5	5	4	8	5
Experience						
Novice	9	5	8	5	52	30
Nonnovice	166	95	139	95	121	70
Education						
STEM major	31	18	35	25	66	38
Ed major	79	46	39	28	67	39
Other major	59	34	64	46	39	22
	M	*SD*	*M*	*SD*	*M*	*SD*
Content knowledge	2.65	0.29	2.51	0.32	2.86	0.33
Implementation						
Total	2.65	0.29	2.51	0.31	2.86	0.33
Contrasting cases	2.40	0.59	2.28	0.58	2.84	0.45
Spaced testing	2.78	0.35	2.63	0.344	2.85	0.37
Visualization	2.76	0.39	2.64	0.37	2.89	0.48
Total	184		153		181	

& Muthén, 1998–2012). To account for the different means across content units, all teacher implementation scores, content knowledge scores, and student achievement scores were standardized by unit prior to aggregation.

Interview Data

In analyzing the interview data, we followed the procedures outlined by Miles and Huberman (1994), Patton (1990), and Coffey and Atkinson (1996). Our conceptual framework (see Figure 2), research questions, and relevant literature served as the basis for our initial coding framework for interview transcripts (Alexander, 2001). We then added more themes and subthemes as called for by our ongoing analysis of the transcript data. We used the constant comparative method to develop the codes (Glaser & Strauss, 1967) so that ideas from the transcripts were used to expand and refine the coding system. Through this iterative process, we changed, adapted, and integrated categories or themes (Goetz & LeCompte, 1984). In this way, we were able to interactively identify themes using both our conceptual framework and the transcript data. This method enabled us to use the data to inductively test our hypotheses as well as to

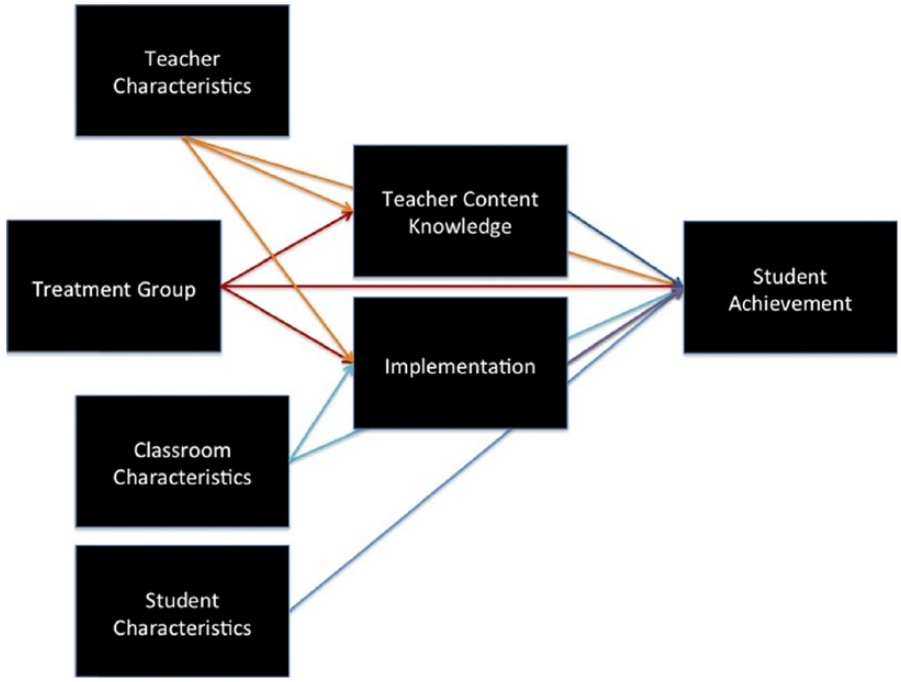

FIGURE 2. *Conceptual model.*

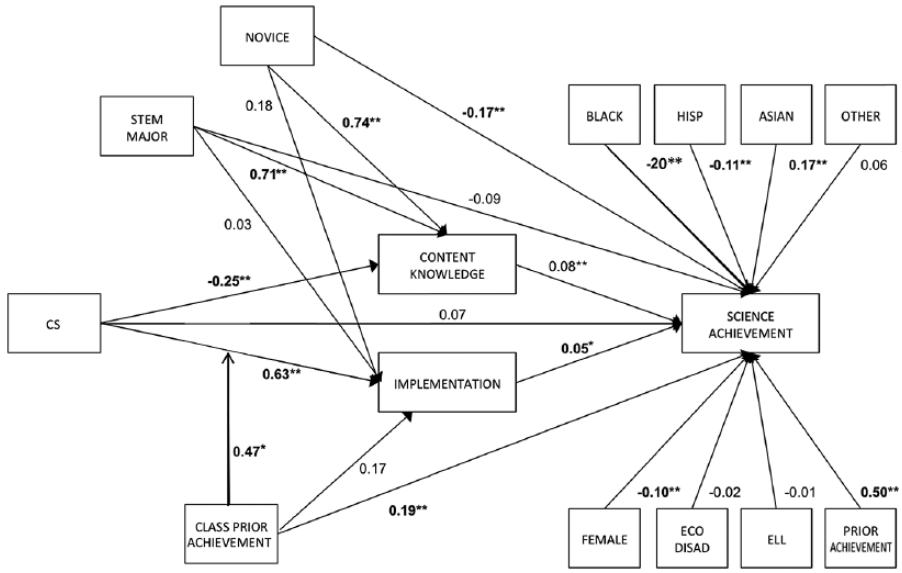

FIGURE 3. *Fitted structural equation model with unstandardized coefficients.*
Note. CS = cognitive science; ELL = English language learners; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error approximation.
*Significant at .1 level. **Significant at .05 level.

deductively allow themes and explanations that we had not anticipated in our conceptual framework to emerge (Emerson, Fretz, & Shaw, 1995; Green, Dixon, & Zaharlock, 2002).

Specifically, the codes we used included, but were not limited to, the following categories: perception of content knowledge in science; experience with science teaching; confidence in science; description of learning in the PD and changes in instruction; understanding of the PD goals; understanding of the CS principles; self-assessment of student science learning, engagement, and motivation; relevant school or classroom context/conditions; interacting with other teachers; and teacher autonomy. In the reporting of results, we illustrate themes with key quotes as exemplars (see Atkinson, Coffey, & Delamont, 2003), a technique that, according to Ryan and Bernard (2003), is "a widely used method for describing themes . . . that lead the reader to understand quickly what it may have taken the researcher months or years to figure out" (p. 282). We included analysis that was relevant to our research questions—including teacher responses that were illustrative and consistent with the results of the survey analysis, but also responses that provide alternative views. We focused on interview data that were directly relevant to our research questions; thus, any interview data omitted from our analysis we viewed as ancillary to the focus of this study. Given the modest size of our sample, we thought it prudent to try to reflect the continuum of teacher experiences rather than focus only on those issues that "most" teachers discussed. Overall, however, most key themes were reflected in the majority of teacher interviews.

Results

Our discussion below focuses on results from the structural equation model assessing the mediating effects of content knowledge and total implementation (a composite measure of implementation of the three principles of CS emphasized in the intervention) on student achievement. The fitted model is detailed in Figure 3. Table 3 provides the results. Owing to our large sample size, model chi-square is not a reliable fit statistic; however, global goodness-of-fit measures indicate a good model fit (comparative fit index [CFI] = 0.873, Tucker–Lewis index [TLI] = 0.719, root mean square error approximation [RMSEA] = 0.019).

As shown in Figure 3, the treatment condition had a direct but nonsignificant effect on student achievement and a significant direct effect on both content knowledge and implementation. Classroom prior achievement moderated the effect of treatment condition on implementation. Content knowledge and implementation were additionally both found to have significant, positive relationships with student achievement. Lending support to our basic model, our covariates were significant predictors of student achievement in ways consistent with previous research. Being a novice teacher had a significant negative relationship with student achievement ($\beta = -.17$, $p = .005$). Analysis of student-level characteristics showed that Black ($\beta = -.20$, $p = .000$), Hispanic ($\beta = -.11$, $p = .01$), and female ($\beta = -.10$, $p = .000$) students had significantly lower achievement than their peers; Asian students ($\beta = .17$, $p = .000$) had higher achievement, and prior achievement ($\beta = .50$, $p = .000$) was significantly positively related to student achievement. These findings suggest the importance of future work that explores whether the intervention worked less well for females and minority students, or, alternatively, suggests further inquiry into factors unaccounted for in the model that could explain gender and race disparities.

Research Question 1: To what extent do classroom implementation of the intervention and teacher content knowledge mediate the intervention's effects on student achievement?

As expected, being assigned to the CS intervention is positively associated with teachers' implementation of the principles of CS, such that teachers in the CS condition report levels of overall implementation 0.63 standard deviations higher than teachers in the control condition, even after controlling for major and years of experience ($p = .000$). Implementation of the principles of CS in turn is positively related to student achievement, after controlling for other student, classroom, and teacher characteristics. Indeed, for every one standard deviation increase in frequency of implementing the CS principles, there is a corresponding 0.05 standard deviation increase in student achievement ($p = .06$). See Figure 3 for details.

Contrary to expectations, we find that participation in the CS intervention is negatively

TABLE 3
Model Output

	Overall	Cells	IRE	ITM
	β	β	β	β
Cognitive science				
Content knowledge (T)	**−0.246**†	**−0.363**†	−0.258	−0.112
Implementation (T)	**0.628**†	**0.686**†	**0.739**†	**0.503**†
Student achievement	0.074	0.099	−0.055	0.113
Content knowledge				
Novice (T)	**0.735**†	**0.789**†	**0.577**†	**0.829**†
Stem major (T)	**0.705**†	**0.634**†	**0.509**†	**0.946**†
Implementation				
Novice (T)	0.178	0.535	0.080	0.127
Stem major (T)	0.029	−0.343	0.072	0.118
Class prior achievement	0.165	−0.133	0.053	0.381
Class prior achievement (moderator)	**0.465***	0.788	0.452	0.293
Student achievement				
Content knowledge (T)	**0.075**†	0.039	**0.098**†	**0.090**†
Implementation (T)	**0.052***	0.094	**0.065**†	0.028
Female (S)	**−0.099**†	**−0.071**†	**−0.113**†	**−0.105**†
Black (S)	**−0.199**†	**−0.131**†	**−0.174**†	**−0.274**†
Hispanic (S)	**−0.106**†	**−0.129***	−0.064	**−0.113***
Asian (S)	**0.166**†	**0.177**†	**0.142**†	**0.157**†
Other (S)	0.060	0.156	0.055	−0.024
Economically disadvantaged (S)	−0.022	−0.017	0.008	−0.069
ELL (S)	−0.005	0.020	0.045	0.011
Prior achievement (S)	**0.501**†	**0.503**†	**0.504**†	**0.498**†
Class prior achievement	**0.190**†	**0.266**†	**0.170***	0.136
Novice (T)	**−0.167**†	**−0.197***	−0.083	**−0.203***
Stem major (T)	**−0.087***	−0.123	−0.094	−0.003
Implementation with content knowledge	0.033	0.048	−0.030	0.102
Indirect effects				
Content knowledge (T)	**−0.018***	−0.014	−0.025	−0.010
Implementation (T)	**0.033***	0.064	**0.048***	0.014
Model fit				
CFI	0.873	0.816	0.806	0.910
TLI	0.719	0.593	0.571	0.800
RMSEA	0.019	0.019	0.024	0.024

Note. IRE = Inside the Restless Earth; ITM = Introduction to Matter; S = student characteristic; T = teacher characteristic; ELL = English language learner; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error approximation.
†Significant at $p = .5$ level. *Significant at $p = .1$ level.

associated with teachers' content knowledge. On average, teachers in the CS condition had levels of content knowledge that are 0.25 standard deviations lower than teachers in the control condition ($p = .04$). There are a number of possible explanations for this finding. For

instance, it may be that while the treatment group focused on CS pedagogical strategies, the business-as-usual control condition received content-intensive PD that increased their content knowledge, leading to higher scores on the content knowledge assessments. This would imply that, in the context of a finite PD window, the emphasis on CS principles and pedagogy may lead to trade-offs in curricular enrichment that suppress teachers' growth in content knowledge. Alternatively, this finding may reflect "unhappy random assignment" where, at baseline, the control group had higher content knowledge. We cannot test this hypothesis, as we did not administer a content knowledge pretest. We did not administer pretests because this study was designed as an RCT, which theoretically creates intervention and control groups that are equivalent. Even so, in future studies, we do recommend pretests on key variables like teacher knowledge, given the possibility of lack of baseline equivalence. In addition, this finding emphasizes the importance of detailed tracking of the "treatment" received by a control group, as additional information on "business as usual" PD could shed light on the extent to which teachers received additional content knowledge. Nevertheless, as expected, content knowledge is positively related to student achievement, such that for every one standard deviation increase in content knowledge, there is a corresponding 0.08 standard deviation increase in student achievement ($p = .001$).

Thus, we find that teachers' implementation of CS principles and teacher content knowledge are significant mediators of the effect of the CS intervention on student achievement. In fact, there was no significant direct effect of the CS intervention after inclusion of implementation and content knowledge in our model, suggesting that the effect is fully mediated by these paths. Note that diagnostics indicated that teacher content knowledge and implementation were not significantly related to each other.

Our interview data help to explain these findings, suggesting several mechanisms by which the CS intervention may have had positive effects on student learning: through application of the CS principles, through improved classroom management, and through collaborative discussions that elevated practice. Also, contrary to our survey findings, several teachers did report increases in their content knowledge.

Application of CS Principles

Interview data revealed that teachers were indeed changing their practice in ways envisioned by the CS intervention. For example, Betty said that the intervention's focus on using students' background knowledge fostered a major difference in her instruction. Whereas previously she would focus on memorizing scientific facts, "[the intervention] is not memorization . . . it's using their background knowledge and it's . . . finding what they already know . . ." Another teacher, Ryan, explained that

> . . . the biggest change [in my instruction] would be starting off with the comparing and contrasting activities, you know doing the back to front, where I show 'em what it looks like first and then go to explain . . . I . . . didn't [previously] approach science that way.

Chloe indicated that the intervention also affected her instruction. Specifically, she reported,

> It [the intervention] changed the way I thought about the way students learn . . . So I found myself more interested in how are they gonna be able to relate things together and what kinds of things to do with the students to make sure that they can build relationships and understanding . . . I found that it enhanced ways to deliver the science content to the students.

Classroom Management and More Efficient Lesson Planning

Extending what we could learn from the survey data, interviews revealed that intervention effects might be partly explained by the fact that for many teachers, use of the intervention materials improved classroom structure and organization and made lesson planning more efficient, thus allowing for more learning time. As Ryan described,

> To have . . . a set formula helps with just the overall management of the lesson and class cause they know what to expect is coming . . . they know I'm gonna start with a warm-up, they know we're gonna move into probably some visualizations next and then they know that we're gonna . . . go over the book or do a

> lab or something else . . . it just . . . definitely has a nice pattern . . . which . . . is helpful . . . in managing . . . the class and managing the lesson.

Jack's comments were consistent with Ryan's. Jack indicated that the structure of the intervention was

> a way of providing a rhythm to the class and [it] . . . set the tone . . . of . . . how the class is going to be approached, so I think that stabilized and . . . that's an important way to keep . . . things managed.

Teachers also reported that the *Casebook* helped with pacing and streamlining their lesson planning, thereby allowing them more time for improving the quality of their instruction. For example, Ryan indicated that "[the *Casebook*] saves me time on lesson planning . . . it follows the curriculum so well . . . it helps me stay on pace." Jack offered that

> . . . [the lessons are] already established . . . so there's never gonna be a flop day [so] . . . there's more of a chance for me to improve the overall course of the lessons each of the days because I have these materials available.

Professional Community: Discussion About Trial and Error

Our interviews suggested that participation in the PLCs, a component of the PD provided by the intervention, influenced the quality of implementation by providing an opportunity for teachers to share their experiences in building their expertise, in using the intervention materials, and in applying the CS principles. Chloe remarked that the PLC discussions were "helpful to [discuss] what kinds of things worked, what didn't work, [and] how people maybe increase the success they had in their individualized classrooms." Specifically, the PLCs afforded opportunity for teachers with different levels of experience and content knowledge to share ideas with each other. As Jack explained,

> The peers in the PLC were the most helpful of and the resources, the sharing of ways to present things, the communal like being able to, to discuss . . . generally there's, there's a good mix of different types of teachers that have different experience levels and so we're all sharing our, our unique ideas about how to present the material and we're all learning from each other.

Other teachers described similar experiences, also highlighting the opportunities that the PLCs created to learn from and get support from other teachers.

Increases in Content Knowledge

Although our survey analysis detected small negative effects between the intervention and content knowledge, in interviews, several teachers indicated that they believed the intervention PD bolstered their science knowledge considerably. This suggests that there may be other explanations for the difference in content knowledge between treatment and control, such as baseline differences or control ("business-as-usual") PD that more heavily focused on content, and in doing so emphasizes the value of using qualitative analysis to supplement and explain quantitative findings. For instance, Betty, who was trained as an elementary and special education teacher and therefore self-identified as having room to grow in her knowledge about science, explained,

> Now that I'm doing this I think like I have an even better understanding of [science]. Like I feel like I can stand in front of the class and not have to look at my book to say, you know this is this, this is this, I can just say to them, like I'm teaching them instead of reading it . . . I have such a better understanding.

Lisa also commented on how the PD improved her content knowledge:

> I mean that helped my personal knowledge much better. I mean I feel like I'm much more familiar with rocks and the rock cycle than I was before the PLC, or the summer development. I think in general it helped, I mean it did nothing but help my, my knowledge of science for sure . . . I mean I knew the 3 types of rocks and that's pretty much it . . . it gave me a lot more knowledge and so when my kids ask questions I feel a little, I feel better about answering them than I did before.

Similarly, Gina said that "[the PD] made me more comfortable. It made me more comfortable with what I was teaching. That I feel like I, I have a, a true knowledge of what it is I'm teaching . . ."

Research Question 2: How, and in what ways, is teachers' implementation of the intervention influenced by teacher experience,

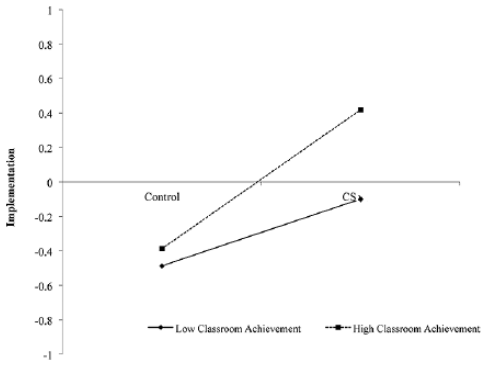

FIGURE 4. *Interaction between implementation and classroom achievement.*
Note. CS = cognitive science.

subject expertise (as measured by college major), and prior classroom achievement?

Teacher experience and STEM major had significant positive associations with teacher content knowledge (β =.74, p = .000; β =.71, p = .000, respectively), but neither had a significant association with implementation of the intervention. Classroom prior achievement was not significantly related to implementation but was a significant moderator of the relationship between treatment condition and implementation (β = .47, p = .067). The size of the gap in implementation scores between low- and high-achieving classrooms varied by treatment condition. We define high and low achievement as one standard deviation above and below the mean, respectively. As shown in Figure 4, in the control group, teachers had roughly equivalent low levels of implementation regardless of classroom prior achievement (−0.49 vs. −0.39). However, in the CS condition, teachers with high-achieving classes had substantially higher levels of implementation than teachers with low-achieving classes (−0.10 vs. 0.42). Classroom prior achievement also had a statistically significant association with student achievement (β = .19, p = .001), such that students in higher achieving classes performed better than students in lower achieving classes, even after controlling for their own prior achievement.

Data from teacher interviews suggest several reasons why implementation might be less robust in lower achieving classrooms, including challenges to adapting the intervention to meet the needs of lower achieving students, having to spend more time on test preparation, and having shorter science periods.

Challenges to Adapting the Intervention for Lower Achieving Students

One reason the intervention may have been implemented with more fidelity in higher achieving classrooms was that teachers indicated the need to adapt it for lower achieving students—one such adaptation included covering less content. Chloe remarked that it was critical to learn from other teachers "what did they do with the Special Ed. students especially with the quizzes and some of the work that was really higher level when we had the lower level . . . readers." Other teachers reported that they had to veer off the planned lessons to provide students with background information or review due to students' lack of prior science knowledge. For example, Ryan said,

> My girls didn't get science last year. Their science periods were . . . watching movies or doing worksheets, just . . . goofing off. So they . . . had no science background . . . I kept saying . . . things like how long our species has been around . . . then all of a sudden, they're like, well where did our species come from? So . . . I just had to stop [and review] . . .

Testing Preparation

Another possible explanation for why implementation was not as rigorous in low-achieving classrooms is that in many of those classes, teachers were more likely to be required to spend additional time on test preparation, which takes away from other instructional science time. Lisa explained,

> Well my class period is an hour and 15 minutes, but the first, I have to do corrective math during my science work . . . so it's anywhere from 45 minutes to a half hour for science . . . my [science] classes were cut because to do the math and now we're doing [State Standardized Assessment] review which I need to do in science for math and . . . that . . . cuts into that time.

Similarly, Betty said that "I really have to move on, I have to keep going" through the material quickly, to allow time for test preparation. Lisa echoed this sentiment: "My classes were cut . . . to do the math and now we're doing [State

Standardized Assessment] review which I need to do in science for math and that . . . cuts into that time." Conversely, Chloe, a teacher at a higher achieving school reported: "We don't take preparations break here . . . The students have every class every day."

Time

Although not included in our quantitative analyses, interview data revealed that the length of instructional periods likely influenced the quality and quantity of implementation; in particular, teacher reports suggested that lower achieving classrooms were less likely to have long science periods. Jack described the advantages of having longer lesson blocks:

> The benefit that I had versus other teachers from talking to them is that I had 75 minutes of instruction time for each lesson, so I had more time to develop a, go one step further in a lot of . . . days, whereas . . . the other teachers . . . have like 2 or 3 days a week for like 15 minutes or like half an hour or something small, a small amount of time, and there's some consistency in having the curriculum, the, the exposure consistently and for a long period of time.

Gina indicated that she had to move faster through the material with shorter class periods:

> Some of us get to teach 90 minutes, some of us get to teach 45, so there's a time constraint. This year I don't have the privilege of 90 minutes. I teach both 8th grades in a 45 minute period of time . . . So I'm a little bit behind maybe what somebody else might be . . . I am just moving at a fast clip.

Betty summed it up by saying "time always seems to be an issue and when I go to the PLCs, time [to fully implement the lessons] is an issue with a lot of people."

Discussion

As education research continues to advance our understanding of effective educational practices through the use of high-quality, rigorously designed randomized studies, there is a greater need to more fully understand why and how some reforms are able to produce effects whereas others fail to do so (Cordray & Pion, 2006). Within and across interventions, there is considerable variation in the extent to which teachers increase their knowledge, implement new practices with high fidelity, understand instructional changes, and elicit effects on student achievement (e.g., Domitrovich et al., 2009; Garet et al., 2010; Gowlett et al., 2015; Penuel et al., 2007).

Our study of the implementation of a middle school science intervention was designed to analyze the role of implementation in eliciting effects, and to better understand how teacher expertise and classroom context may serve as barriers and facilitators to implementation. Our findings contribute to the field's growing understanding of teachers' implementation of new interventions, including how, why, and under what circumstances they work to change instruction in ways that foster improvements in student learning.

Results from the study should be interpreted with several caveats. The interview data are suggestive, not conclusive. Furthermore, our structural equation modeling was outside the bounds of the RCT. A more robust test of the role of previous student achievement, for example, would be to randomly assign teachers of low-achieving classrooms to intervention and control groups, and examine effects. RCTs rarely have the resources to randomize on such factors, so we use quasiexperimental techniques instead. Furthermore, we recognize that teacher content knowledge is complex and is not fully represented by undergraduate major, a content knowledge test, or any other single measure. Although we view including both survey and interview data on implementation as a strength, we also readily acknowledge that observations of implementation are able to detect aspects of quality that surveys and interviews are not (Desimone, 2009).

Implementation Matters, and Ought to Be Measured

We found that teachers were implementing the CS principles, and that implementation had direct effects on student achievement, which is consistent with research supporting the use of the CS principles. These findings highlight the importance of measuring implementation. Although research has demonstrated that high levels of implementation fidelity translate into improved student outcomes (Durlak, 2010; Durlak & DuPre, 2008; Kaderavek & Justice, 2010; Stein et al., 2008), in some cases, teachers

may not be implementing the intervention due to contextual or environmental pressures, lack of knowledge, or any one of a myriad of other factors. *Before concluding that an intervention does not have effects on student learning because RCT results show no effects, it is imperative that researchers have a measure of whether or not the teachers actually implemented the intervention as intended.* Ideally, as implementation studies become more sophisticated, we will be able to establish thresholds, of both frequency and quality, at which effects are elicited, and target particular dimensions of an intervention that are most effective. Such results would be especially valuable for shaping policy.

An Intervention May Work Partly Through Secondary Mechanisms

Perhaps more importantly for the education policy implementation field in general, our findings suggest that an instructional intervention may work through mechanisms secondary to the approach being taught in the intervention PD. *Teachers reported that their use of the intervention's structured and sequenced activities improved their classroom management, and provided a more coherent organizational structure to their daily lessons*. These pedagogical routines were secondary to the CS principles being taught in the PD, yet it was the instructional routines teachers focused on in reporting how the intervention changed their approach to teaching. This is consistent with previous research showing teachers may focus their learning more on the pedagogical strategies used in a PD, rather than the content focus (Covay Minor, Desimone, Caines, & Hochberg, 2016). This raises important questions to consider in designing, implementing, and evaluating classroom interventions. How closely aligned should a PD or intervention be to curricular materials, such as pacing guides, to help make the link between content and instructional approaches? Recent rigorous RCTs have shown that PD that is explicitly linked to the curriculum or pacing guide are much more likely to be adopted (Fishman, Marx, Best, & Tal, 2003; Roschelle et al., 2010), compared with those that rely strictly on building teachers' content knowledge or introducing strategies that teachers then have to decide for themselves when and how to integrate into classroom lessons (e.g., Garet et al., 2008).

An Intervention's Effectiveness May Be Related to Balancing Teacher Content Knowledge, Aligned Lesson Plans, and Teacher Invention

There has been considerable attention paid to understanding the complex dimensions of teacher content knowledge (Schulman, 1986), and how to improve it in ways that elicit more student learning (Sadler et al., 2013; Smith, Neergard, Hochberg, & Desimone, 2017). In our study, teachers who had STEM majors and more experience teaching scored higher on our content knowledge test, and these higher scores predicted higher student achievement scores. These findings are consistent with previous work linking teacher content knowledge and student learning (Hill et al., 2005; Metzler & Woessmann, 2012). However, *we found that there was no significant relationship between content knowledge and implementation, and furthermore that teachers with higher content knowledge did not implement the intervention more frequently or better as evidenced by the fact that the CS intervention had a* negative *effect on content knowledge.* This is consistent with previous working showing new teachers' content knowledge was not linked to their growth in instructional quality (Desimone, Hochberg, & McMaken, 2016).

This raises a question about trade-offs between a scripted intervention and one that requires considerable teacher knowledge and invention. In our study, teachers indicated that the structured, sequential nature of the lesson plans that accompanied the intervention helped improve the management and organization of their teaching, allowing for more student-learning time and fewer "flop" lessons. Although some scholars criticize scripted interventions for taking away teacher creativity (Richardson & Placier, 2001), others have found that interventions requiring considerable teacher input are unlikely to be implemented well (in large part due to time constraints), and thus see fewer good effects on students as compared with more scripted interventions (Correnti & Rowan, 2007).

What is an appropriate balance between "scripted" lessons—which provide detailed pacing, lessons, and activities—and more conceptual interventions that rely exclusively on

teachers building their content knowledge (e.g., they are not linked to the teacher's day-to-day lessons)? We interpret our findings, in light of related literature, to support a hybrid approach—a balance between providing teachers with materials such as daily lesson activities aligned to the curriculum *and* PD that supports knowledge building related to the use of the new strategies and/or content. Furthermore, the daily lessons should allow room for teacher invention and creativity. For example, because our intervention relied on active learning strategies such as discussion and applications, there was considerable opportunity for teachers to apply their own insights and creativity to these activities. *We suspect that the success of the intervention was partly due to the balance of research-based approaches (i.e., applying CS principles to teaching), PD that included both content and pedagogy, and implementation that provided aligned-lesson guidance while still allowing for teacher creativity and invention.* We recommend that the development and evaluation of similar classroom interventions pay attention to the interaction of these important factors.

PLCs Help Teachers Refine and Adapt an Intervention

Another mechanism through which the intervention seemed to work was the PLC discussions that were part of the intervention PD. *Teachers indicated that PLCs were valuable opportunities to discuss trial-and-error efforts, to find out what was working, and to share ideas and experiences with other teachers to improve their implementation of the CS principles.* Although the role of professional communities as a mechanism of teacher learning is well established (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; J. W. Little, 1982), not as much is known about their role in implementation fidelity. How should PLCs be used in shaping and refining implementation of an intervention? How salient are they in establishing quality, and in developing and sanctioning constructive adaptations? How should they be organized to foster high-quality implementation? In our study, PLCs were especially important in refining strategies, and we suspect that these findings would apply across contexts.

Another important use of the PLCs, as reported in teacher interviews, was to discuss how to adapt the intervention to low-achieving students, large classes, interference with required test preparation, and so on. This is consistent with our survey analysis finding that implementation within the CS condition was better in higher achieving classrooms. Lower implementation in low-achieving classrooms is especially problematic, given that many interventions are designed to help address the achievement gap. We know that a particular model of instruction may influence students differently depending on students' achievement trajectories and other background factors (Desimone & Long, 2010), and that practitioners routinely adapt practice to match classroom context and meet the needs of their students (McHugo et al., 2007), with varying success on improving student outcomes (Durlak, 2010). Our finding here supports the notion that *ideas for adaptation should be integrated into intervention PD. Intervention/PD designers would do well to think through and discuss how teachers can best implement their intervention given at least the most frequently documented challenges in lower achieving schools, such as shorter class periods (e.g., for subjects besides math or reading), more test prep time, and lack of student background knowledge. Furthermore, PLCs provide valuable forums for teachers to discuss strategies for adaptation.* Follow-up PD might even include coaching on how to adapt to particular student needs. Alternatively, designers might consider these common challenges during the development stage and create interventions that explicitly take into account shorter class periods, lack of student background knowledge, and other limitations that are common in lower achieving schools.

In terms of implications for evaluating interventions, we recommend analyzing not only how implementation varies by subpopulations but also whether the impact of implementation on outcomes varies by key subpopulations (Harn, Parisi, & Stoolmiller, 2013).

Conclusion

Our study provides insights relevant for shaping intervention PD, monitoring classroom implementation, and guiding implementation

analyses. We found that the frequency of implementation helped to explain effects found in an RCT of a middle school science intervention. Integrating findings from teacher interviews with a structural equation model that analyzed achievement and survey data, we found that the CS intervention may work not only through applying principles of CS but also through fostering better classroom management and organization and developing professional communities, and we suggest that these may be common phenomena across similar interventions that provide sequenced pedagogical strategies or activities and opportunities for collaborative discussions that elevate practice. Furthermore, our results have implications for informing decisions about how to balance a focus on increasing teacher content knowledge, on one hand, and providing explicit pedagogical strategies linked to the curriculum, on the other. We additionally found that for teachers participating in the intervention, classrooms with lower prior achievement had lower implementation scores, which interview data suggest is due to adaptations that slowed teachers down, and having less science class time due to testing preparation. Our study highlights the importance of anticipating and calibrating interventions to the contextual complexities of real-life classrooms, and identifies several factors with the potential to contribute to improved design and evaluation of such interventions.

Appendix

Survey Items

Compare and Contrast Survey Items

Q30. When teaching compare and contrast exercises to your TARGET SECTION, how often did you ask students to . . .

Scale: Never, Sometimes, Often, Always

- Complete compare and contrast exercises before introducing the topic and general concepts for a lesson sequence
- Complete compare and contrast exercises at the end of a lesson sequence

Q31. When teaching compare and contrast exercises in the Holt Cells, Heredity, and Classification unit to your TARGET SECTION, to what extent did you ask students to . . .

Scale: Never, Sometimes, Often, Always

- List several features that are common across examples
- List several features that are different across examples

Q32. When teaching compare and contrast exercises in the Holt Cells, Heredity, and Classification unit to your TARGET SECTION, to what extent did you select compare and contrast examples to . . .

Scale: Never, Sometimes, Often, Always

- Highlight features shared by all members of a category
- Highlight features that distinguish categories from one another
- Show common/typical examples of categories

Q33. When teaching the Holt Cells, Heredity, and Classification unit in your TARGET SECTION, to what extent did you select compare and contrast examples to . . .

Scale: Never, Sometimes, Often, Always

- Foster student mastery of broad scientific principles (big ideas) of the chapter

Q34. When teaching compare and contrast exercises during the Holt Cells, Heredity, and Classification unit in your TARGET SECTION, to what extent were the exercises . . .

Scale: Never, Sometimes, Often, Always

- Teacher- and student-led—I first model some similarities and differences for students, then ask the students to highlight similarities and differences

Q35. When teaching compare and contrast exercises during the Holt Cells, Heredity,

and Classification unit in your TARGET SECTION, to what extent did you . . .

Scale: Never, Sometimes, Often, Always

- Present the relevant information from each exercise for students
- Ask students to do research to find the relevant information

Testing Survey Items

Q36. In teaching the Holt Cells, Heredity, and Classification unit to your TARGET SECTION, how often did you . . .

Scale: Never, Once during the unit, Two or three times during the unit, More than three times during the unit

- Give students a test or quiz in class
- Test each of the broad scientific principles (big ideas) of the unit

Q37. Across all of the tests and quizzes that students completed during the Holt Cells, Heredity, and Classification unit, which of the following is true for your TARGET SECTION? (Check one answer for each row.)

Scale (answer categories):
None of the broad scientific principles (big ideas) in the unit
Some of the broad scientific principles (big ideas) in the unit
Most of the broad scientific principles (big ideas) in the unit
All of the broad scientific principles (big ideas) in the unit

- Students were asked to study . . .
- The combined pool of questions I gave touched upon . . .

Q38. In teaching the Holt Cells, Heredity, and Classification unit to your TARGET SECTION, how often did you . . .
Scale (answer categories):
I never did this
For some tests/quizzes
For most or all tests/quizzes

- Have students engage in organized review activities before taking a quiz/test
- Indicate which items students got right or wrong on the test or quiz
- Engage in class discussion focused on problem areas identified in a quiz/test

Q39. Considering all the tests and quizzes you gave for the Holt Cells, Heredity, and Classification unit in your TARGET SECTION, to what extent did you . . .

Scale (answer categories):
I never did this
For some tests/quizzes
For most or all tests/quizzes

- Use an entire quiz/test provided by Holt
- Use quiz/test questions not provided by Holt
- Test content from only one chapter at a time
- Test content from prior chapters in the unit

Visualization Survey Items

Q40. During your teaching of the Holt Cells, Heredity and Classification unit in your TARGET SECTION, when an image in the textbook used arrows, color coding, or a cut-away perspective, to what extent did you . . .

Scale (answer categories):
Never
For a few of the relevant images in the Cells, Heredity, and Classification Unit
For some of the relevant images in the Cells, Heredity, and Classification Unit
For most or all relevant images in the Cells, Heredity, and Classification unit

- Discuss the image's arrows only if students expressed confusion about the image
- Discuss the image's arrows regardless of whether students expressed confusion about the image
- Discuss the image, but not the arrows
- Discuss the use of color coding as it relates to that particular image
- Discuss the use of color coding as it relates to images in a general way

- Discuss the use of a cut-away perspective only when it was central to understanding the image's main idea
- Discuss the use of cut-away perspective separate from the image's main idea
- Discuss the image, but not the cut-away perspective

Authors' Note

The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grants R305B090015 and R305A120288 to the University of Pennsylvania and University of California, Los Angeles (UCLA).

Note

1. In our original preliminary analyses, we created dummy variables to distinguish math, science, and other majors from each other. We found no significant differences among types of science majors. Because this is a science-focused intervention and we had a limited sample size, we combined majors as keeping these distinctions limited our degrees of freedom.

References

Alexander, R. (2001). *Culture and pedagogy: International comparisons in primary education*. Malden, MA: Blackwell.

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*, 87–113.

Allison, P. D. (2001). *Missing data* (*Vol. 136*). SAGE. Retrieved from http://statisticalhorizons.com/wp-content/uploads/2012/01/Milsap-Allison.pdf

Atkinson, P., Coffey, A., & Delamont, S. (2003). *Key themes in qualitative research: Continuities and change*. Walnut Creek, CA: AltaMira Press.

Becker, B. J., & Aloe, A. M. (2008, March). *Teacher science knowledge and student science achievement*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Berends, M. (2000). Teacher-reported effects of New American School designs: Exploring relationships to teacher background and school context. *Educational Evaluation and Policy Analysis*, *22*, 65–82.

Berends, M., Chun, J., Ikemoto, G. S., Stockly, S., & Briggs, R. J. (2002). *Challenges of conflicting school reforms: Effects of New American Schools in a high-poverty district*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monograph_reports/MR1483.html

Berman, P., & McLaughlin, M. W. (1976). Implementation of educational innovation. *The Educational Forum*, *40*, 345–370.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., Roitman, D. B., & Emshoff, J. G. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology*, *15*, 253–268.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness*, *1*, 237–264.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*, 61–100.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: The University of Chicago Press.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*, 369–378.

Chi, M. T. H. (2005). Common sense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences*, *14*, 161–199.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145–182.

Chi, M. T. H., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement:

Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*, 673–682.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data analysis: Complementary strategies*. Thousand Oaks, CA: SAGE.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*, 175–199.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. Bootzin & P. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103–124). Washington, DC: American Psychological Association.

Correnti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Education Research Journal*, *44*, 298–337.

Council of Chief State School Officers. (2005). *Surveys of enacted curriculum: Tools and services to assist educators*. Washington, DC: Author.

Covay Minor, E., Desimone, L. M., Caines, J., & Hochberg, E. (2016). Insights on how to shape teacher learning policy: The role of teacher content knowledge in explaining differential effects of professional development. *Education Policy Analysis Archives*, *24*(38). Retrieved from http://epaa.asu.edu/ojs/article/view/2365

Cromley, J. G., Perez, A. C., Fitzhugh, S., Tanaka, J., Newcombe, N., Shipley, T. F., & Wills T. W. (2010). Improving students' diagrammatic reasoning: A classroom intervention study with eye tracking data. Paper presented at the annual conference of the American Educational Research Association, Denver, CO.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45.

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

Desimone, L. M., & Garet, M. S. (2015). Best practices in teacher's professional development in the United States. *Psychology, Society, and Education*, *7*, 252–263.

Desimone, L. M., Hochberg, E., & McMaken, J. (2016). Teacher knowledge and instruction in beginning teaching: How do they grow and how are they linked? *Teachers College Record*, *118*(5), 1–54.

Desimone, L. M., McMaken, J., & Cherng, S. (2010). *Conceptualizing and Measuring Implementation of Education Interventions*. Paper presented at the Annual Meeting of the American Educational Research Association, 2010. Denver, CO.

Desimone, L. M., & Long, D. (2010). Does conceptual instruction and time spent on mathematics decrease the student achievement gap in early elementary school? Findings from the Early Childhood Longitudinal Study (ECLS). *Teachers College Record*, *112*, 3024–3073.

Desimone, L. M., Richards, M., & Hwang, J. (2013, May 21–22).Implementation of a cognitive science intervention. *Schools of the 21st Century Conference*, Washington, DC.

Diamond, B. S., Maerten-Rivera, J., Rohrer, R., & Lee, O. (2014). Effectiveness of a curricular and professional development intervention at improving elementary teachers' science content knowledge and student achievement: Year 1 results. *Journal of Research in Science Teaching*, *51*, 635–658. doi:10.1002/tea.21148

Domitrovich, C. E., Gest, S. D., Gill, S., Bierman, K. L., Welsh, J. A., & Jones, D. (2009). Fostering high-quality teaching with an enriched curriculum and professional development support: The Head Start REDI program. *American Educational Research Journal*, *46*, 567–597. doi:10.3102/0002831208328089

Duncan, G. J., & Magnuson, K. A. (2003). The promise of random-assignment social experiments for understanding well-being and behavior. *Current Sociology*, *51*, 529–541.

Durlak, J. A. (2010). The importance of doing well in whatever you do: A commentary on the special section, "implementation research in early childhood education." *Early Childhood Research Quarterly*, *25*, 348–357. doi:10.1016/j.ecresq.2010.03.003

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, *18*, 237–256.

Emerson, R., Fretz, R., & Shaw, L. (1995). *Writing ethnographic fieldnotes*. Chicago, IL: University of Chicago Press. doi:10.7208/chicago/9780226206851.001.0001

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430–457.

Fairchild, A. J., MacKinnon, D. P., Taborga, M. P., & Taylor, A. B. (2009). R 2 effect-size measures for

mediation analysis. *Behavior Research Methods*, *41*, 486–498.

Feiman-Nemser, S. (2012). *Teachers as learners*. Cambridge, MA: Harvard Education Press.

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

Fleer, M. (2009). Supporting scientific conceptual consciousness or learning in "a roundabout way" in play-based contexts. *International Journal of Science Education*, *31*, 1069–1089. doi:10.1080/09500690801953161

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

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., . . . Silverberg, M. (2008). *The impact of two professional development interventions on early reading instruction and achievement*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pdf/20084030.pdf

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

Garet, M. S., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., . . . Doolittle, F. (2010). *Middle school mathematics professional development impact study: Findings after the first year of implementation* (NCEE 2010-4009). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20104009/pdf/20104010.pdf

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*, 393–408.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Goetz, J., & LeCompte, M. (1984). *Ethnography and qualitative design in educational research*. Orlando, FL: Academic Press.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on education productivity. *Journal of Human Resources*, *32*, 505–523. doi:10.2307/146181

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*, 129–145. doi:10.3102/01623737022002129

Gowlett, C., Keddie, A., Mills, M., Renshaw, P., Christie, P., Geelan, D., & Monk, S. (2015). Using Butler to understand multiplicity and variability of policy reception. *Journal of Education Policy*, *30*, 149–164. doi:0.1080/02680939.2014.920924

Green, J., Dixon, C., & Zaharlock, A. (2002). Ethnography as a logic of inquiry. In J. Flood, J. Jensen, D. Lapp, & J. Squire (Eds.), *Handbook for methods of research on English language arts teaching* (pp. 201–224). Mahwah, NJ: Lawrence Erlbaum.

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children*, *79*, 181–193.

Harris, D. N., & Sass, T. R. (2007). *Teacher training, teacher quality, and student achievement* (Working Paper No. 3). Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*, 85–110.

Hegarty, M., Kriz, S., & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition & Instruction*, *21*, 325–360.

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

Holme, J. J., & Rangel, V. S. (2012). Putting school reform in its place: Social geography, organizational social capital, and school performance. *American Educational Research Journal*, *49*, 257–283.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*, 88–110.

Jüttner, M., Boone, W., Park, S., & Neuhaus, B. J. (2013). Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). *Educational Assessment, Evaluation and Accountability*, *25*, 45–67. doi:10.1007/s10763-012-9384-6

Kaderavek, J. N., & Justice, L. M. (2010). Fidelity: An essential component of evidence-based practice in speech-language pathology. *American Journal of Speech-Language Pathology*, *19*, 369–379. doi:10.1044/1058-0360(2010/09-0097)

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). London, England: Guilford Press.

Kurtz, K. J., Miao, C. H., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences*, *10*, 417–446.

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, *38*, 635–652.

Little, J. W. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal*, *19*, 325–340. doi:10.3102/00028312019003325

Little, T. D., Card, N. A., Bovaird, J. A., Preacher, K. J., & Crandall, C. S. (2007). Structural equation modeling of mediation and moderation with contextual factors. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 207–230). Mahwah, NJ: Psychology Press.

Loeb, S., Beteille, T., & Kalogrides, D. (2012). Effective schools: Teacher hiring, assignment, development, and retention. *Education Finance and Policy*, *7*, 269–304.

Luft, J. A., Dubois, S. L., Nixon, R. S., & Campbell, B. K. (2015). Supporting newly hired teachers of science: Attaining teacher professional standards. *Studies in Science Education*, *51*, 1–48.

Massey, C., Cleland, D., & Mandel, B. (2013, September). *Professional development interventions in a large-scale randomized controlled study of middle school science learning*. Invited symposium presented at the Fall 2013. Conference of the Society for Research on Educational Effectiveness, Washington, DC.

McHugo, G. J., Drake, R. E., Whitley, R., Bond, G. R., Campbell, K., Rapp, C. A., . . . Finnerty, M. T. (2007). Fidelity outcomes in the national implementing evidence-based practices project. *Psychiatric Services*, *58*, 1279–1284. doi:10.1176/appi.ps.58.10.1279

McLaughlin, M. W. (2005). Listening and learning from the field: Tales of policy implementation and situated practice. In A. Lieberman (Ed.), *The roots of educational change* (pp. 58–72). Dordrecht, The Netherlands: Springer Netherlands.

Metzler, J., & Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics*, *99*, 486–496.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: SAGE.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, *11*, 247–266.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

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

Newcombe, N. S. (2013). Seeing relationships: Using spatial thinking to teach science, mathematics, and social studies. *American Educator*, *37*(1), 26–40.

Nowicki, B. L., Sullivan-Watts, B., Shim, M. K., Young, B., & Pockalny, R. (2013). Factors influencing science content accuracy in elementary inquiry science lessons. *Research in Science Education*, *43*, 1135–1154. doi:10.1007/s11165-012-9303-4

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*, 33–84.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: SAGE.

Penuel, W. R., Fishman, B., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, *44*, 921–958. doi:10.3102/0002831207308221

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, *36*, 399–416.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal*, *48*, 965–995.

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

Porter, A. C., Polikoff, M. S., Barghaus, K. M., & Yang, R. (2013). Constructing aligned assessments using automated test construction. *Educational Researcher*, *42*, 415–423. doi:10.3102/0013189X13503038

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*, 25–31.

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

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., . . . Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*, 833–878.

Rowan, B., Chiang, F., & Miller, R. J. (1997). Using research on employees' performance to study the effects on students' achievement. *Sociology of Education*, *70*, 256–284. doi:10.2307/2673267

Ruiz-Primo, M. A. (2005, April). A multi-method and multi-source approach for studying the fidelity of implementation. In S. Lynch (Chair) & C. L. O'Donnell, *"Fidelity of implementation" in implementation and scale-up research designs: Applications form four studies of innovation science curriculum materials and diverse populations*. Symposium conducted at the annual meeting of the American Educational Research Association, Montreal, Québec, Canada.

Ryan, G. W., & Bernard, H. R. (2003). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Collecting and interpreting qualitative materials* (2nd ed., pp. 259–309). Thousand Oaks, CA: SAGE.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, *50*, 1020–1049.

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

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, *16*, 475–522.

Scull, J., Porter, A. C., Merlino, J., & Massey, C. (2017). *Infusing cognitive science principles into middle school science materials and teaching methods*. University of Pennsylvania, Philadelphia, PA.

Shallcross, T., Spink, E., Stephenson, P., & Warwick, P. (2002). How primary trainee teachers perceive the development of their own scientific knowledge: Links between confidence, content and competence? *International Journal of Science Education*, *24*, 1293–1312. doi:10.1080/09500690110110106

Smith, T., Neergard, L., Hochberg, E., & Desimone, L. M. (2017). Organizational effects on teacher quality. *Teachers College Record*, in press.

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Sáenz, L., Yen, L., & Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis*, *30*, 368–388.

Summerfelt, W. T. (2003). Program strength and fidelity to evaluation. *Applied Developmental Science*, *7*, 55–61.

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

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2011). Intervention fidelity in special and general education research journals. *The Journal of Special Education*, *47*, 3–13.

U.S. Department of Education. (2015). *What Works Clearinghouse: Demystifying the What Works Clearinghouse: A webinar for developers and researchers*. Washington, DC: Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/document.aspx?sid=246&pid=5#fidelity

Webster-Stratton, C., Reinke, W. W., Herman, K. C., & Newcomer, L. L. (2011). The incredible years teacher classroom management training: The methods and principles that support fidelity of training and delivery. *School Psychology Review*, *40*, 509–529.

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

Winters, T. M., Wise, L. L., & Towne, L. (Eds.). (2004). *Advancing scientific research in education*. Washington, DC: National Academies Press.

Yang, R., Porter, A., Massey, C., & Merlino, J. (2017). *Applying cognitive science principles to a middle school curriculum: Comparing two approaches to increasing student achievement* (A version of this paper was presented at the annual meeting of the American Educational Research Association in 2014, Philadelphia, PA).

Yoon, K. W., Duncan, T., Wen-Yu Lee, S., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (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.

Authors

LAURA M. DESIMONE's research focuses on the effects of education policy at the state, local, and classroom levels. She is especially interested in studying school and classroom implementation, and the effects of teacher learning interventions on teachers and students.

KIRSTEN LEE HILL is an advocate for cross-stakeholder collaboration and strives to make research relevant and accessible to, and informed by, all stakeholders in public education.

Manuscript received August 12, 2016
First revision received November 29, 2016
Second revision received January 27, 2017
Accepted January 29, 2017