Workshop on Education
Tue Nov 3, 2020
100 pm until 220 pm

Heather C. Hill
Jerome T. Murphy Professor of Education
Harvard Graduate School of Education

Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis

ABSTRACT: I will present results from a meta-analysis of 95 experimental and quasi-experimental preK-12 science, technology, engineering, and mathematics (STEM) professional development and curriculum programs, seeking to understand what content, activities and formats relate to stronger student outcomes. Across rigorously conducted studies, we found an average weighted impact estimate of +0.21 standard deviations. Programs saw stronger outcomes when they helped teachers learn to use curriculum materials; focused on improving teachers' content knowledge, pedagogical content knowledge and/or understanding of how students learn; incorporated summer workshops; and included teacher meetings to troubleshoot and discuss classroom implementation. I will also show models that suggest the impacts of instructional improvement programs operate through changes in practice, rather than changes in teacher knowledge. I will conclude by discussing implications for policy and practice.

BIO: Heather C. Hill is the Jerome T. Murphy Professor of Education at the Harvard Graduate School of Education. Her primary work focuses on teacher and teaching quality and efforts to improve both. She has also developed instruments for measuring teachers' mathematical knowledge for teaching (MKT) and the mathematical quality of instruction (MQI) within classrooms. With Susanna Loeb, she writes the *What Works, What Doesn't* column for Education Week. Her other academic interests include teacher professional development, research design, and policy implementation. She is the coauthor, with David K. Cohen, of *Learning Policy: When State Education Reform Works* (Yale Press, 2001).

Join Zoom Meeting
https://uchicago.zoom.us/j/96154177840?pwd=VkNJOFdwSDlxTU12cVgxdzhJZytlZz09

Meeting ID: 961 5417 7840
Password: 823149

Educational Evaluation and Policy Analysis
Month 201X, Vol. XX, No. X, pp. 1–34
DOI: 10.3102/0162373719849044
Article reuse guidelines:sagepub.com/journals-permissions

Strengthening the Research Base That Informs STEM Instructional Improvement Efforts: A Meta-Analysis

Kathleen Lynch
Annenberg Institute at Brown University
Heather C. Hill
Kathryn E. Gonzalez
Cynthia Pollard
Harvard Graduate School of Education

We present results from a meta-analysis of 95 experimental and quasi-experimental pre-K–12 science, technology, engineering, and mathematics (STEM) professional development and curriculum programs, seeking to understand what content, activities, and formats relate to stronger student outcomes. Across rigorously conducted studies, we found an average weighted impact estimate of +0.21 standard deviations. Programs saw stronger outcomes when they helped teachers learn to use curriculum materials; focused on improving teachers' content knowledge, pedagogical content knowledge, and/or understanding of how students learn; incorporated summer workshops; and included teacher meetings to troubleshoot and discuss classroom implementation. We discuss implications for policy and practice.

Keywords: *curriculum, instructional practices, mathematics education, professional development, science education, meta-analysis*

Instructional improvement efforts constitute a persistent feature of the pre-K–12 science, technology, engineering, and mathematics (STEM) educational landscape, with a decades-long trail of new curriculum materials and professional development programs aimed at changing how teachers interact with students around content. Such initiatives bring with them significant costs; scholars estimate that districts spend between 1% and 6% of their budgets on professional development (see, for example, Corcoran, 1995; Miles, Odden, Fermanich, & Archibald, 2004; Miller, Lord, & Dorney, 1994), and the market for instructional materials totaled US$11.9 billion in 2013 (Cavanagh, 2015). Prior to 2002, scholars rarely rigorously evaluated such instructional improvement programs, often using instead cross-sectional and/or self-report data to identify "best practices" in professional development and curricular design (e.g., Becker & Park, 2011; Desimone, 2009; Sparks, 2002). However, following calls in the early 2000s for stronger research into the impact of educational interventions (Confrey & Stohl, 2004; Raudenbush, 2008; Shavelson & Towne, 2001), federal research portfolios began to prioritize research methods that allow causal inference and to use student outcomes as the major indicator of program success.

Dollars' and scholars' turn toward using causal methods and student-level impacts has resulted in a wealth of new studies. These new studies, we argue, permit rigorous empirical analyses linking program characteristics to outcomes, a topic long of interest to practitioners who make decisions regarding intervention

design and/or adoption. In this article, we present results from a meta-analysis of pre-K–12 STEM curriculum materials and professional development programs intended to improve instructional quality and student learning, seeking to understand what content, activities, and formats are linked to stronger student outcomes. Our work differs from other recent efforts in that it is a formal meta-analysis rather than a structured review (e.g., Gersten, Taylor, Keys, Rolfhus, & Newman-Gonchar, 2014; Kennedy, 2016), and because the newly available studies allow us to compile a data set larger than past reviews and, thus, to exclude studies with weaker designs.

We argue that this work is particularly timely. The *Every Student Succeeds Act* requires that districts receiving Title I funds must adopt "evidence-based interventions," including programs and strategies proven to be effective in raising student achievement. However, recent null findings from large-scale studies (e.g., Borman, Cotner, Lee, Boydston, & Lanehart, 2009; Santagata, Kersting, Givvin, & Stigler, 2010), as well as an analysis of studies curated by Malouf and Taymans (2016; What Works Clearinghouse [WWC], 2010), have led many to doubt the efficacy of instructional improvement programs. By contrast, we find an average weighted impact estimate of +0.21 standard deviations among the programs we examine. We describe our study in more detail below.

Background

STEM Instructional Improvement

Recent calls for reform in STEM education have focused on increasing both student understanding of core disciplinary ideas and student engagement with key disciplinary practices such as inquiry, argumentation, and proof (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010; National Research Council [NRC], 2011, 2013). Yet, observational studies have shown that instruction in U.S. STEM classrooms tends to be lacking in disciplinary concepts and low in student cognitive demand (e.g., Banilower, Smith, Weiss, & Pasley, 2006; Hiebert et al., 2005; H. C. Hill, Litke, & Lynch, 2018; Kane, Kerr, & Pianta, 2014).

Perhaps as a result, programs aimed at improving STEM instructional quality abound. These programs tend to involve two main strategies for improvement: teacher professional development, which is typically intended to change some aspect of teachers' instruction, and new curriculum materials, which are typically intended to shape both instruction and the subject matter content teachers teach. Professional development and curriculum programs can be independent (e.g., professional development alone) or used in combination (e.g., when curriculum materials' implementation is supported by professional development). In contrast to alternative reforms, such as standards, test-based accountability, and market-based reforms, which set out broad principles and incentives to which schools and teachers must interpret and react, curriculum and professional development provide direct instructional guidance, attempting to shape schools' and teachers' day-to-day interactions with students through lesson plans and instructional strategies (Ball & Cohen, 1996). We discuss the specific theory of action for both professional development and curriculum below.

Providers of STEM teacher professional development—often districts, but sometimes nonprofits, professional associations, or university-based faculty—typically design a set of experiences intended to affect change in a variety of teacher- and classroom-level phenomena. Most programs focus on instruction as a primary target for change, providing teachers new routines, instructional strategies, and/or ways of teaching content. However, many of these programs also hope to change teachers' beliefs, knowledge of content, and knowledge of how students learn content to support instructional improvement. For instance, Garet et al. (2010) describe a program in which teachers learn rational number concepts, learn about persistent student misconceptions with this topic, and then receive group and individual coaching on how to apply the material learned in the content-focused sessions in their own classrooms. Schneider and Meyer (2012) describe a program in which teachers learn about assessments and formative assessment practices, and then put them into practice. The STeLLA program (Taylor et al., 2015) not only focuses mainly on improving teachers' capacity to conduct analyses of instruction but

also includes sessions on science content and how students learn content. Such programs reflect the view that teachers' instructional practice cannot change without corresponding changes in teachers' beliefs and knowledge (Smith & O'Day, 1990).

As these examples imply, the experiences designed by professional development providers can vary substantially. As critiques of the "one-shot workshop" surfaced two decades ago (Goldenberg & Gallimore, 1991; Kennedy, 1999), providers experimented with new formats, including program delivery distributed across one or more school years, 1:1 coaching, collaborative workgroups, and teacher learning in online settings. Recent reform movements have also changed professional development foci, from programs primarily targeted toward improving teacher content knowledge (Frechtling, Sharp, Carey, & Vaden-Kiernan, 1995) to more diverse topics, including improving teachers' pedagogical content knowledge (Shulman, 1986), helping teachers develop and implement content-specific formative assessment, helping teachers address the needs of English learners, and helping teachers use technology more effectively in STEM classrooms. With these new foci, professional development activities also changed, with newer programs offering more study of student work, more focus on the curriculum materials to be used in classrooms, and more focus on lesson analysis and planning.

The theory of action around curriculum materials also involves shaping what teachers do in classrooms, but here the focus is on both introducing specific instructional moves—supported by the activities and guidance in the text itself—and changing the content taught in substantial ways. For instance, the elementary curriculum materials supported by the National Science Foundation (NSF) in the 1990s were designed not only to engage students in mathematical practices, such as problem-solving, mathematical reasoning, and precise communication, but also to include topics that had not previously been taught in elementary schools, such as data, statistics, and early algebra (Stein, Remillard, & Smith, 2007). In science, numerous projects have built units and curricula intended to replace conventional textbooks with more inquiry-oriented instruction and content that is either more relevant to students' lives or future careers (e.g., Marx et al., 2004; Schwartz-Bloom & Halpin, 2003).

It is worth noting that many providers of curriculum materials also hold a theory of action similar to that of professional developers—that teacher beliefs and knowledge must change to support improvements in classroom practice. In some programs, providers intentionally mix the new curriculum materials with intensive professional development. For instance, Saxe, Gearhardt, and Nasir (2001) paired a new fractions unit with a 5-day summer workshop and 13 in-school sessions focused on improving teachers' knowledge of fractions, teachers' knowledge of how students learn fractions, and teachers' knowledge of student motivation. Roschelle et al. (2010) and Hand, Norton-Meier, Gunel, and Akkus (2016) followed a similar approach. Many curriculum providers also embed guidance for teachers in the written materials themselves. This may include explanations of content intended for teachers—for instance, pages explaining the conceptual ideas within the unit, how they connect to one another, how to represent those ideas, alternative solution methods to problems, and even sample dialogue or scripts. Examples of such curricula include Investigations in Number, Data and Space in mathematics (Russell et al., 2006) and Promoting Science Among English Language Learners (P-SELL; Llosa et al., 2016) in science.

In other cases, curriculum providers offer less professional development, often only a few days intended to orient teachers to routines in the curriculum, adaptations for students with specific needs, and technological enhancements to the curriculum (e.g., Pane, Griffin, McCaffrey, & Karam, 2014; Resendez & Azin, 2006). A small number of curriculum materials providers (e.g., Cervetti, Barber, Dorph, Pearson, & Goldschmidt, 2012) eschew professional development entirely, often in hopes of replicating real-life conditions in schools.

Syntheses of STEM Instructional Improvement Programs

Several recent syntheses examine STEM teacher professional development and curriculum improvement efforts. We review the methodology and findings from these syntheses and then comment on their characteristics.

In the area of professional development (Blank & de las Alas, 2010; Kennedy, 1999, 2016; Scher & O'Reilly, 2009; S. M. Wilson, 2013; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007), prior syntheses have generally indicated positive impacts on learning. Meta-analyses suggest that effect sizes range between +0.21 *SD* (Blank & de las Alas, 2010) and +0.54 *SD* (Yoon et al., 2007) on student outcomes, with some suggestion that content-specific (math) rather than content-general (classroom management) programs produce greater learning gains (Kennedy, 1999; Scher & O'Reilly, 2009). However, Scher and O'Reilly (2009) noted that the pool of STEM-focused articles and reports published by 2004 could not support most expert recommendations regarding "best practices" in professional development (see, for example, Desimone, 2009). For example, although experts have frequently expressed the opinion that professional development must be longer in duration to be effective (e.g., Darling-Hammond & McLaughlin, 1995; Desimone, 2011), recent syntheses of the empirical literature have returned inconsistent findings on this point (e.g., Blank & de las Alas, 2010). Kennedy (2016) found that more time-intensive programs had weaker effects, whereas Yoon found that the three studies in his review that had the least amount of professional development (5–14 hours) showed no statistically significant impacts on student learning. However, Yoon's review included only nine studies. Scher and O'Reilly (2009) found that multiyear interventions were more effective than single-year interventions in math, but not in science.

Two recent structured reviews of professional development, each conducted several years after federal guidelines changed to prioritize causal research, provide another form of evidence regarding instructional improvement programs. One review specific to mathematics (Gersten et al., 2014) used the stringent WWC evidence standards to screen studies, and perhaps as a result returned too few (five) to discern patterns in program impacts. Another (Kennedy, 2016) identified 28 studies, split equally between English language arts (ELA), science, and mathematics. Kennedy found that the characteristics often cited as best practices in professional development (content-focused, collective participation of all teachers within a grade/school, duration) were less predictive of positive outcomes than other features, including assisting teachers in gaining insight into their practice and helping teachers inject novel ideas into their practice. The latter categories included several coaching and curriculum-based programs. Finally, in science, Slavin, Lake, Hanley, and Thurston (2014) found that professional development programs supporting inquiry-based elementary science raised student outcomes by an average of 0.36 *SD*. For secondary science, programs that emphasized inquiry through teacher professional development saw an average effect size of +0.24 *SD*.

In the area of curriculum materials, two research syntheses by Slavin and colleagues (Slavin & Lake, 2008; Slavin, Lake, & Groff, 2009) using studies published between 1971 and 2008 found fairly small effects for mathematics curriculum materials, at +0.03 *SD* and +0.10 *SD* for secondary and elementary curricula, respectively. For science (Slavin et al., 2014), elementary programs with "kits" and accompanying professional development had a near-zero (+0.03 *SD*) average impact on student outcomes. The average impact of kits in secondary science proved similar (+0.04 *SD*), with the average impact of science textbooks not substantially larger (+0.10 *SD*; Cheung, Slavin, Kim, & Lake, 2017). A review of studies by the NRC (2004) conducted in the early 2000s found stronger results for NSF-funded curricula than conventional materials in that 59% of comparisons between materials favored the NSF materials. However, many regard vote counting as a weak and misleading synthesis procedure (Hedges & Olkin, 1980), and Confrey (2006) herself noted that the studies included in the NRC report were of varying quality, and that as study rigor increased, effect sizes diminished. Slavin and colleagues, perhaps because they used more rigorous study selection criteria, did not find any relationship between study design and effect size, although they note that the methodological rigor of most included studies was still relatively weak, and "quality research is particularly lacking in this area" (Slavin, 2008, p. 480).

Finally, one recent study (Taylor et al., 2018) examined interventions that offered professional development, curriculum materials, and computer

software intended to improve student science outcomes. Across 96 such studies, they found an average impact of 0.489 *SD*, with programs evaluated by researcher-designed assessments posting significantly higher impacts, similar to an earlier report by C. J. Hill, Bloom, Black, and Lipsey (2008). Other study and intervention characteristics did not significantly predict program outcomes.

Motivation for the Current Study

These research syntheses share several important characteristics. To start, nearly all noted that the evidence base for making recommendations was thin at the time of the review. Using the pool of articles and reports on professional development published by 2003, Yoon and colleagues (2007) could find only nine ELA, math, and science studies that met WWC evidence standards. Years later, Gersten et al. (2014) located only five mathematics professional development studies that met the WWC's exacting standards. Other evaluations (e.g., Cheung et al., 2017; Scher & O'Reilly, 2009; Slavin et al., 2014) achieved a greater sample size by including quasi-experimental studies with matched comparison groups (and sometimes retrospective matching, a disputed technique; Slavin, 2008). Kennedy's (2016) cross-subject synthetic review with 28 studies was an exception, as many studies described were more recent strong quasi-experiments or actual randomized trials. In general, however, the small sample sizes, typical in these reviews, limited both the generalizability of findings and the number of program features, or moderators, that could be coded and examined. We argue that recent Institute of Education Sciences (IES)- and NSF-funded classroom-level experiments provide a thicker evidence base, eliminating the need to include studies that conducted post hoc matching and facilitating more moderator analyses.

These reviews also share another important characteristic, in that many identified only a small number of program features for study. This is most apparent in the area of professional development, where research syntheses almost universally coded studies for duration, content specificity, and delivery format. For curriculum materials, most analyses categorized programs by the degree of alignment with standards-based reform (e.g., Confrey & Stohl, 2004). With the accumulation of more recent rigorous studies, however, have come opportunities to understand how features of professional development or curriculum materials moderate effects on student outcomes. For instance, professional development activities (watching video, designing lessons, studying student work) may differentially affect program outcomes, as might curriculum design features, such as length of implementation and degree of support for teachers. In addition, the programs studied recently contain more varied delivery methods and features (e.g., coaching, online learning components) than those of a decade ago.

Viewed from a contemporary perspective, several other issues emerge. First, existing research syntheses of professional development studies tend to find that most returned positive and significant effects. Yoon et al. (2007), for instance, found only one of 20 effects identified across nine studies was negative, and only one was zero. Yet, the more recent wave of studies tends to find more mixed impacts (e.g., Garet et al., 2010; Santagata et al., 2010). This suggests that the pool of studies included in earlier syntheses might have suffered from two issues: the "file drawer" problem, in which studies with null results are not published (Slavin, 2008), or the "boutique" problem, in which only highly promising—and often unusual—programs are studied (H. C. Hill, 2004). We believe that both problems may have been ameliorated in recent years. IES and NSF funding guidelines have explicitly encouraged the evaluation of programs that have wide reach, meaning those programs may be more typical of the offerings available to teachers. Furthermore, better reporting practices—for example, final reports posted on websites or short reports archived on conference websites—may have rescued studies from file drawers.

Second, the practice of reviewing professional development and curriculum studies separately creates conceptual and practical difficulties. On the conceptual side, most curriculum programs also include a professional development component for teachers; likewise, some professional development programs offer teachers materials to support the implementation

of new practices within classrooms. On the practical side, the combination of professional development and materials together may be especially effective, as compared with either one alone or one with a minimal dose of the other (Cohen & Hill, 2001). Studying both within one review may enhance our understanding of how these instructional improvement efforts can complement one another.

Method

Search Procedures

We searched for studies in three phases. We began by scanning the reference lists of prior reviews of math and science professional development and curriculum improvement programs (Blank et al., 2008; Cheung et al., 2017; Furtak, Seidel, Iverson, & Briggs, 2012; Gersten et al., 2014; Kennedy, 1999, 2016; Scher & O'Reilly, 2009; Slavin & Lake, 2008; Slavin et al., 2009; Slavin et al., 2014; Timperley, Wilson, Barrar, & Fung, 2008; S. M. Wilson, 2013; Yoon et al., 2007; Zaslow, Tout, Halle, Whittaker, & Lavelle, 2010) for studies published between the years 1989 through 2004. Due to resource constraints, we did not conduct additional searches for materials dated prior to 2004. We argue that this is reasonable, given that prior reviewers have previously searched the published and gray literature from the pre-2004 period extensively.[1] In tandem with the fact that rigorous studies of teacher professional development (PD) and curriculum improvement were relatively rare prior to the early 2000s (e.g., Kennedy, 2016), we considered the likelihood relatively low that we would have uncovered previously unknown randomized trials or sufficiently well-designed quasi-experiments from the pre-2004 period.[2] Supplemental Appendix Table C13 in the online version of the journal compares studies published pre-2004 versus studies published 2004 or later. As illustrated in the table, descriptively, studies from the earlier period had larger effect sizes on average; they were also less likely to be randomized controlled trials (RCTs), and less likely to use a state standardized test as an outcome variable. As we discuss below, we control for these methodological variables in all of our primary models.

In the second search phase, we conducted an electronic search using the databases Academic Search Premier, ERIC, Ed Abstracts, PsycINFO, EconLit, and ProQuest Dissertations & Theses, for the period January 2004 to March 2016. Searches were conducted using subject-related keywords adapted from Yoon et al. (2007) and methodology-related keywords designed to capture experimental and quasi-experimental methods adapted from Kim and Quinn (2012).[3] We also searched the websites of Regional Education Labs, WWC, the World Bank, Inter-American Development Bank, Empirical Education, Mathematica, MDRC, and American Institutes for Research (AIR) for relevant materials. We also searched the abstracts of the Society for Research on Educational Effectiveness (SREE) conference. We ceased our materials search in March 2016.

In the third search phase, we downloaded, from the NSF Community for Advancing Discovery Research in Education (CADRE) and IES websites, a list of all STEM award grantees from the years 2002 to 2012. We then conducted electronic database and Web searches to find all studies published from each award. In the case of 29 awards, we could find no publicly available reports or information that included student outcomes, and we attempted to contact project principal investigators (PIs) via email to obtain study results. Of these 29 grant awards, we were sent, or later located, reports from 17. Of these, the reports from 16 studies were excluded according to the screening criteria below, and one was included in our analyses.

The search procedures described above netted a total of 8,099 records identified through database screening, and an additional 1,391 records identified through other sources (see Figure 1 for screening flowchart). After removing duplicates, we were left with 7,926 records.

Screening Procedures

Screening proceeded in two phases. First two raters screened each of the studies' titles and abstracts to identify potentially relevant studies, passing studies into the second phase when they covered grades pre-K–12, included student outcomes, included quantitative data, and focused on math- and science-specific content and/or instructional strategies. All studies flagged as

potentially relevant by either rater were then reviewed by one of the authors, who made a final determination about moving the study forward. A total of 656 studies met the initial relevance criteria and were advanced to full-text screening.

In the second screening phase, two authors examined the full text of each study and applied more detailed content and methodological criteria. To qualify for inclusion in the final data set, we required that studies use a randomized or quasi-experimental research design with a comparison group. Following Slavin and Lake (2008), we included studies that assembled comparison groups via prospective, but not post hoc, matching. Slavin and Lake have argued that "Prospective studies, in which experimental and control groups were designated in advance and outcomes are likely to be reported whatever they turn out to be, are always to be preferred to post hoc studies, other factors being equal." According to Slavin and Lake, post hoc designs have several characteristic problems. First, when researchers attempt to examine the efficacy of a program by simply comparing the test scores of schools or classrooms that completed the program with those that did not, only "survivors" are included in the treatment group. This can lead to upwardly biased estimates of the treatment impact, if schools or classrooms began the treatment but dropped it because it was not working. Second, when researchers construct post hoc comparisons of treated and matched comparison schools, they generally have many potential "comparison" schools to choose from; this raises the concern that researchers may inadvertently (or even intentionally) choose "comparison" schools that have made less academic progress over the study period than the treatment schools. Third, post hoc matching studies are often commissioned by textbook publishers and curriculum developers because they are low cost and easy to conduct. For these same reasons, study authors may easily abandon them if they do not show positive results, and, thus, the retrievable studies may be especially skewed toward positive findings (Slavin & Lake, 2008). To classify a study as prospective matching, we required that the study authors explicitly report that they matched treatment and comparison groups on pretest data prior to the intervention's commencement. Studies that explicitly stated that matching was done retrospectively, or that were silent on this point, were excluded. We also excluded studies in which treatments had been in place prior to pretesting. We also required that studies present pretest means, and that pretest differences between groups be less than 50% of a standard deviation. Despite this relatively liberal threshold for allowable pretest differences, these data requirements resulted in the inclusion of only nine quasi-experiments.

Studies had to be published in 1989 or later, be written in English, and have participating students in grades pre-K–12. The program had to focus on classroom-level STEM instructional improvement through professional development, curriculum materials, or both. We excluded programs with no instructional improvement component, such as after-school peer tutoring or computerized at-home skills practice, as well as studies that examined the impact of *variables*, rather than programs.[4] Studies had to include sufficient data to calculate an effect size for student outcomes. The most common reasons for exclusion were for characteristics of the intervention (e.g., off-topic, not a classroom-level intervention; $N = 561$), methodology issues (e.g., no control group, post hoc design; $N = 310$), and sample issues (e.g., did not have at least two teachers and 15 students in each condition; $N = 45$). Note that some studies had multiple reasons for exclusion (see Figure 1).

Ninety-five studies met the review inclusion criteria and advanced to the study coding phase. In cases where we encountered multiple versions of the same study, we used all available reports to glean information about the study, and used the most recent version (often the peer-reviewed publication version) for impact estimates. Of the studies that met the inclusion criteria, 44% were from peer-reviewed journal articles; 23% were conference papers or presentations; 18% were technical reports; 5% were district, state, or federal government reports; 4% were doctoral dissertations; and 5% were from other sources. Many of these studies reported multiple effect sizes due to the inclusion of multiple outcome measures, multiple versions of the same program with a common control group, multiple samples, or multiple programs.

The final analysis sample includes 258 effect sizes nested within these 95 studies. This includes

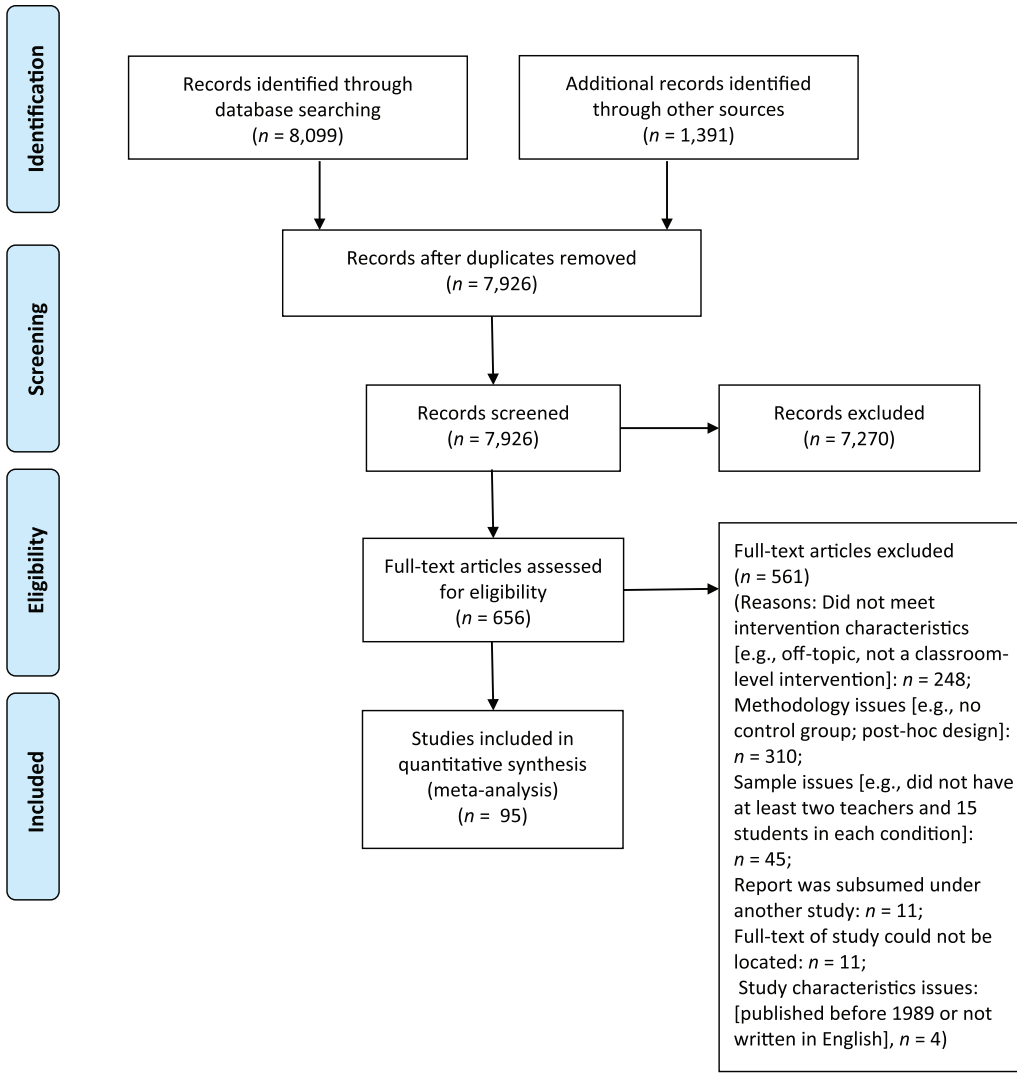

FIGURE 1. *PRISMA study screening flowchart.*
Source. Moher, Liberati, Tetzlaff, Altman, and The PRISMA Group (2009).
Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

a separate effect size for each treatment contrast, each assessment of math or science achievement, and each sample of students and teachers reported by the study.[5] To account for the nested nature of our data, our analysis used the robust variance estimation (RVE) approach (Tanner-Smith & Tipton, 2014), as discussed below.

Study Coding

To develop content codes, we began by examining prior meta-analyses and reviews of instructional improvement interventions, naming both broad categories (professional development, curriculum, technology, subject matter) and specific codes (teachers studied curriculum materials, teachers solved math problems) that commonly appeared in the literature. We then used this skeleton structure to code a sample of studies, modifying existing codes to be more clear, and adding codes as needed to cover additional program features. We also added a set of codes that captured study size, design, and quality. Finally, our technical advisory board, comprising university faculty

and other experts in teacher professional development and curriculum materials, reviewed our codes, making suggestions for amending and adding as necessary.[6]

For programs involving professional development, four primary coding categories emerged. A first pair of codes captured professional development *duration*, indexed as the number of contact hours teachers spent in professional development experiences (both for professional development and curriculum materials programs) and as the time span over which those hours were spread. A second set of codes recorded the main *focus* (or foci) of the professional development, including emphases on improving teacher content knowledge, exploring how students learn, integrating technology into the classroom, and learning how to use curriculum materials. Third, we coded for specific *activities* that teachers engaged in during the professional development, such as observing a demonstration of instruction or working through student curriculum materials. A final set of codes captured the *format* of the professional development, for instance, whether it was delivered during a summer workshop, contained coaching, or involved online learning. Within each of the three substantive coding categories, we also created an indicator of the number of codes applied, hypothesizing that professional development programs with multiple foci, activities, or formats (e.g., a focus on both how students learn and how to use curriculum materials) might be more effective compared with more narrowly focused programs.

For programs that involved new curriculum materials, we coded for whether the curriculum provided implementation guidance (e.g., text describing possible student–teacher dialogues around content) or supplied teachers with kits for implementation. We also recorded the proportion of a lesson the curriculum was intended to replace (i.e., 1%–100%), and the total number of minutes the curriculum was intended for use in a school year.

We also designed codes to capture the rigor of the design as well as aspects of its implementation. Specifically, to record potential selection bias of participants into groups, we coded whether treatment and control groups were formed via random assignment or by prospectively matching individuals/classrooms/schools on baseline data. To explore potential impacts of general and differential attrition bias, we included two measures of potential effect size-level attrition problems: (a) author-reported attrition of more than 20% at the student or cluster level and (b) differential attrition between treatment and control groups of more than 10%. We note, however, that our use of binary indicators rather than raw attrition rates constitutes a limitation, as continuous attrition measures would have allowed for more precise estimation of the impacts of attrition. Finally, we coded whether authors simply reported a given outcome as "nonsignificant" to capture the potential for bias due to selective reporting.

We note we were unable to capture other potential sources of bias such as performance and detection bias, or biases stemming from participants' and researchers' knowledge of whether participants were in the treatment or control group. Given that it is evident to participants and researchers alike whether teachers are engaged in new professional development and curriculum programs, in most cases it is not possible to blind participants and researchers to condition.

We also coded for a variety of other study descriptors, such as publication type (peer-reviewed vs. not) and whether the study took place in the United States or abroad. We examine whether any of these study descriptors relate to effect sizes. Based on evidence that effect sizes vary by type of test (C. J. Hill et al., 2008; Taylor et al., 2018), we also coded for the nature of the student assessment, including state or district standardized tests (e.g., the Texas Essential Knowledge and Skills assessment), standardized tests available through commercial vendors (e.g., Woodcock–Johnson), and researcher-designed assessments. We expected the last to be more sensitive to the content of instructional improvement programs, and, thus, potentially provide larger effect sizes.

Coding of full-text studies was conducted by the study authors, along with two trained research assistants. Before beginning operational coding, we went through a process of establishing interrater reliability. Each week, each member of the team coded two studies, then met to reconcile disagreements and refine the code descriptions. This process continued until coders reached 80% agreement (4 weeks

for low-inference codes such as bibliographic information, 5 weeks for higher inference codes describing program characteristics). Once we achieved a stable set of codes and the 80% threshold, each study was coded by two researchers, including at least one study author, working as a pair. Each researcher coded studies independently, and met to reconcile discrepant records. All disagreements were resolved through discussion.

Effect Size Calculation

We calculated standardized mean difference effect sizes using Hedges's *g*:

$$g = J \times \frac{\left(\overline{Y_E} - \overline{Y_C}\right)}{S^*}.$$

In this formula, $\overline{Y_E}$ represents the average treatment group outcome, $\overline{Y_C}$ represents the average control group outcome, and S^* represents the pooled within-group standard deviation. J is a correction factor that adjusts the standardized mean difference to avoid bias in small samples:

$$J = 1 - \frac{3}{4 \times \left(N_E + N_C - 2\right) - 1}.$$

In this formula, N_E represents the number of students in the treatment group and N_C represents the number of students in the control group. Effect sizes were calculated using the software package *Comprehensive Meta Analysis* for the majority of cases. We used the following decision rules to calculate effect sizes: If the authors reported a standardized mean difference effect size, such as Cohen's *d* or Glass's delta, we converted author-reported effect sizes to Hedges's *g* (52% of effect sizes).[7] If authors did not report a standardized mean difference effect size but did report a covariate-adjusted unstandardized mean difference (e.g., a coefficient from a multilevel model) and raw standard deviations, we calculated a standardized mean difference effect size and converted to Hedges's *g* (15%). If covariate-adjusted mean differences were not reported, we calculated effect sizes based on raw posttest means and standard deviations (22%). If neither standardized mean differences nor raw means and standard deviations were reported, effect sizes were calculated based on the coefficients and standard errors of multilevel or regression models (4%).[8] In the remaining cases, effect sizes were calculated from other results (e.g., studies that reported the results of analyses of variance [ANOVAs]; 7%).

We applied corrections to study-reported standard errors when necessary to account for the clustering of students within classrooms or schools (Littell, Corcoran, & Pillai, 2008).[9] For the five studies that did not report the number of clusters, we followed Scher and O'Reilly's (2009) procedure for imputing that number based on available information, typically estimating the total number of classrooms from the total number of students. When intraclass correlations could not be inferred from the study report, we also followed Scher and O'Reilly (2009) in adopting the WWC-recommended default intraclass correlations of .20.

We had 29 outcomes where the study did not provide enough information to calculate an effect size.[10] All came from studies that did report an effect size for at least one additional achievement outcome, thus no studies were dropped from the analysis due to missing outcomes. Furthermore, many of these missing outcomes were subscales, meaning they were a smaller selection of items already represented in the reported effects. Therefore, we ignore these missing outcomes in the main analyses. However, we also examine the sensitivity of our results to the inclusion of these outcomes by imputing a range of values for these missing effect sizes and reestimating our primary models.[11]

Model Selection

A common problem in meta-analyses occurs when single studies yield multiple effect sizes, as authors often measure more than one outcome. These nested effect sizes are likely to be correlated, violating the assumption of statistical independence. To address this issue, prior meta-analyses in this area have averaged effect sizes (Kennedy, 2016; Scher & O'Reilly, 2009; Slavin et al., 2014; Yoon et al., 2007). Here, we argue that an RVE approach developed by Tanner-Smith and Tipton (2014) more properly

models our data. This approach adjusts standard errors to account for the dependencies between effect sizes, and is analogous to methods for adjusting standard errors in ordinary least squares (OLS) regression models for heteroscedasticity (e.g., using Huber–White standard errors) or to account for the nesting of data within clusters (e.g., clustered standard errors). Importantly, this approach allows for the inclusion of multiple effect sizes from the same study within the meta analysis, and, therefore, avoids the loss of information that would occur by either dropping effect sizes or including a single average effect size for each study (for a more detailed discussion, see Tanner-Smith & Tipton, 2014). RVE models can also control for methodological features known to affect outcomes (e.g., C. J. Hill et al., 2008; Taylor et al., 2018). This approach has been used in several recent meta-analyses where individual studies report multiple effect sizes (e.g., Clark, Tanner-Smith, & Killingsworth, 2016; Dietrichson, Bøg, Filges, & Klint Jørgensen, 2017; Gardella, Fisher, & Teurbe-Tolon, 2017).

The RVE approach is designed to account for two types of dependencies among effect sizes. The first type of dependency is *correlated effects*, which arises when a study provides multiple effect size estimates for a single underlying construct or of correlated underlying measures, or when the same control group is used for multiple treatment contrasts. The second type of dependency is *hierarchical effects*, which occurs when multiple treatment–control contrasts are nested within a larger cluster of experiments (e.g., a research group conducts multiple evaluations of the same program). When both of these dependencies are present, Tanner-Smith and Tipton (2014) recommend selecting a method based on the more frequent type of dependence. Because correlated effects predominated in our data, we used the recommended inverse variance weights recommended by Tanner-Smith and Tipton (2014).

The weight for effect size i in study j is calculated by the following:

$$w_{ij} = \frac{1}{\left\{\left(v_{*j} + \tau^2\right)\left[1 + \left(k_j - 1\right)\rho\right]\right\}},$$

where v_{*j} is the mean of within-study sampling variances (SE_{ij}^2) within each study, τ^2 is the estimate of the between-studies variance component, k_j is the number of effect sizes within each study, and ρ is the assumed correlation between all pairs of effect sizes within each study. Therefore, effect sizes from studies with more effect sizes and higher sampling variances are given lower weight. It is assumed that ρ is constant across studies, and we use the recommended default value of $\rho = .80$ (Tanner-Smith & Tipton, 2014). However, simulation studies suggest that results using the RVE approach are not sensitive to values of ρ (e.g., Tanner-Smith & Tipton, 2014; S. J. Wilson, Tanner-Smith, Lipsey, Steinka-Fry, & Morrison, 2011). We also conduct a series of sensitivity checks to determine whether our results are sensitive to alternative values of ρ. We use the *robumeta* package in Stata 15 (developed by Tanner-Smith & Tipton, 2014) to estimate our RVE models and incorporate the small-sample correction proposed by the developers of the RVE approach (Tanner-Smith & Tipton, 2014; Tipton & Pustejovsky, 2015). We also report the results of F tests to test the joint significance of the program features included in our RVE models. These F tests were conducted using the *robumeta* and *clubSandwich* packages in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).

As noted above, we grouped potential moderators of program impact into five categories indicating the *time*/*duration, focus, activities*, and *format* of professional development, and the characteristics of new curriculum materials. To examine whether specific features in each category moderated program impact, we fit five sets of conditional meta-regression models with RVE, including the coded features as moderators and treating these moderators as fixed. Within each category, we first modeled the effect of each code separately. To further understand their joint relationships, we then fit a model entering all codes together. All models controlled for whether the study was an RCT, whether the effect size estimate controlled for covariates, the type of student assessment used, whether the program focused on mathematics (rather than science), and an indicator variable for whether the study was conducted at the preschool level.

In some cases, there is within-study variability in program features (e.g., among studies with multiple treatment arms). Following the

recommendation of Tanner-Smith and Tipton (2014), we included the study-level mean value of each covariate and moderator. For the two covariates where there is within-study variability in at least 10% of studies (state standardized test, other standardized test), we also included a within-study version of the covariate that is calculated by subtracting the study-level mean values from the original covariate values. The full data set is available from the authors on request.

Finally, we note that the RVE method addresses heterogeneity of effect sizes differently compared with traditional meta-analyses. As described by the RVE developers, the primary aim of this method is to estimate fixed effects, such as meta-regression coefficients, rather than to model variation in effect sizes. As a result, tests for heterogeneity used in traditional meta-analysis are not available with the RVE approach (Tanner-Smith & Tipton, 2014; Tanner-Smith, Tipton, & Polanin, 2016). However, we do report the method-of-moments estimate of τ^2 for each of our primary models as measures of between-study heterogeneity in effect sizes.

Results

Descriptives and Overall Average Impacts

Table 1 provides descriptive statistics regarding the study designs and programs included in our data set. Of the included studies, 22% included a treatment arm that focused on professional development alone (without the introduction of new curriculum materials), 9% included a treatment arm that focused on curriculum materials alone (without the provision of professional development), and 75% included a treatment arm that focused on both new curriculum materials and professional development. Most included studies had randomized designs; only nine quasi-experimental studies met the criteria for inclusion. Roughly two thirds of studies focused on mathematics rather than science. Table 1 also shows study-level frequencies for the format, timing and duration, foci, and activities of programs with any professional development component, and for characteristics of curriculum materials programs.[12] Because we had a relatively small sample size, when two variables correlated at the .40 level or above, we combined them into a single predictor indicating whether the program had either feature. This occurred in two cases (a focus on content knowledge/pedagogical content knowledge and knowledge of how students learn, and under activities, solving problems, and working through student materials).

Across all included studies, we found an average weighted impact estimate of +0.21 standard deviations (see Table 2). To contextualize the magnitude of this effect, a typical treatment group student would be expected to rank about 8 percentile points higher than a typical control group student (Lipsey et al., 2012). We found a method-of-moments estimate of the between-study variance in study average effect sizes of 0.037, with a between-study standard deviation of 0.191, indicating that most studies had small to moderate positive impacts. Of the 258 effect sizes included in the 95 identified studies, 214 effect sizes (83%) were positive in sign and 88 effect sizes (34%) were statistically significant. Only 40 effect sizes were negative in sign (16%) and only four were negative and statistically significant (2%). Four effect sizes had point estimates of zero (2%). The unweighted average effect size for math outcomes was +0.27 *SD* and the average effect size for science outcomes was +0.18 *SD*. However, the results of conditional meta-regression model estimated using RVE indicate no statistically significant difference in mean effect sizes based on whether programs focused on math or science (see Table 3). We, therefore, included both math and science outcomes in all analyses.

As shown in Table 3, the type of assessment employed was a strong predictor of effect size magnitude, with impacts for both standardized (e.g., Woodcock–Johnson) and state-standardized tests lower by about 0.27 *SD* ($p < .01$) relative to impacts for researcher-designed assessments. Table 3 also shows that there were not statistically significant differences in effect sizes based on study design (i.e., whether the study was an RCT), whether study authors adjusted for covariates such as student demographic indicators, whether the study sample included preschool students, and subject matter.

Table 4 shows that the average effect size for programs that incorporated both professional development and new curriculum materials was approximately 0.10 *SD* ($p < .05$) larger than the

TABLE 1

Categories and Descriptions of Codes

Code	Code description	Code present[a]
Intervention type		
Professional development only	Study included only professional development	22%
Professional development and curriculum materials	Study included both professional development and new curriculum materials	75%
Curriculum materials only	Study included only new curriculum materials	9%
Research design and sample characteristics		
RCT	Study is an RCT	91%
Subject matter—Math	Subject matter focus of study is math or math/science	64%
Preschool	Study sample included preschool students	19%
Effect size type		
State standardized test	Outcome is a state standardized test	17%
Other standardized test	Outcome is from other standardized test	30%
Adjusted for covariates	Effect size is adjusted for covariates (e.g., pretest score)	76%
PD format[b]		
Same-school collaboration	Teachers participated in professional development with other teachers from their own school	74%
Implementation meetings	Teachers met formally or informally with other activity participants to discuss classroom implementation (e.g., troubleshooting meeting)	35%
Online professional development	Part or all of the professional development was conducted online	18%
Summer workshop	The professional development included a summer workshop	54%
Expert coaching	The professional development involved coaching or mentoring from experts who observed instruction and provided feedback (e.g., a debriefing meeting, via video or live)	20%
PD lead by researchers/intervention developers	The PD was led by the intervention developers and/or the study authors	64%
PD timing and duration[c]		
Contact hours	Total number of PD contact hours	45 hours
Time span over which professional development was conducted		
Less than 1 week	The PD was conducted over less than 1 week	15%
One week	The PD was conducted over 1 week	3%
One month (8–30 days)	The PD was conducted over 8–30 days	3%

(continued)

TABLE 1 (CONTINUED)

Code	Code description	Code present[a]
One semester (31 days to 4 months)	The PD was conducted over 31 days to 4 months	13%
One year	The PD was conducted over 4 months to 1 year	49%
More than 1 year	The PD was conducted over more than 1 year	16%
PD focus[b]		
Generic instructional strategies	The professional development was focused on content-generic instructional strategies (e.g., improving classroom climate and student motivation)	11%
How to use curriculum materials	The PD focused on how to use curriculum materials	75%
Integrate technology	The PD focused on how to integrate technology into the classroom	11%
Content-specific formative assessment	The PD focused on formative assessment strategies specific to mathematics and science teaching (e.g., strategies to elicit student understanding of fractions or the scientific method)	17%
Improve content knowledge/pedagogical content knowledge/how students learn	The PD focused on improving teachers' pedagogical content knowledge (e.g., how students learn mathematics or science)	55%
PD activities[b]		
Review sample student work	Teachers studied examples of students' work (including watching videos of students)	16%
Observed demonstration	Teachers observed a video or live demonstration/modeling of instruction	35%
Solved problems/worked through student materials	Teachers solved problems or exercises during the PD or worked through student materials during the PD	42%
Developed curriculum/lesson plans	Teachers developed curricula or lesson plans during the PD	19%
Reviewed own student work	Teachers studied examples of their own students' work during the PD	11%
Curriculum materials[d]		
Implementation guidance	The curriculum materials provided teachers with implementation guidance (e.g., support for student–teacher dialogues around the content)	43%
Laboratory/hands-on experience or curriculum kits	The curriculum materials included materials/guidance that supported inquiry-oriented explorations (e.g., science laboratory or hands-on mathematics kits)	28%
Curriculum dosage (hours)	Total number of hours that the curriculum was intended to be used	66.6 hours
Curriculum proportion replaced (%)	The proportion of each lesson that the new curriculum was intended to replace existing curriculum	91%

Note. $N = 95$ studies. RCT = randomized controlled trial; PD = professional development.

[a]Figures in third column include percentage of studies that feature the row code for binary variables or the sample average calculated at the study level for continuous variables (e.g., contact hours and curriculum dosage). For studies that had the feature present in one treatment arm but not another treatment arm, the code is counted as present if it is present in any treatment arm (e.g., a study with one treatment arm including only curriculum materials and a second treatment arm including both PD and curriculum materials would be included in both rows).

[b]Codes for PD focus, activities, and format were counted as "Not Present" for studies that did not involve a PD component.

[c]PD timing/duration excludes studies without a PD component.

[d]Codes for features of interventions involving new curriculum materials were counted as "Not Present" for studies that did not involve new curriculum materials.

TABLE 2

Results of Estimating an Unconditional Meta-Regression Model With RVE

	Dependent variable: Effect size (Hedges's *g*)
Constant	0.209***
	(0.025)
N effect sizes	258
N studies	95
τ^{2} [a]	0.037
95% prediction interval[b]	[−0.165, 0.583]

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. RVE = robust variance estimation.
[a]τ^2 is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).
[b]The 95% prediction interval is calculated as the estimated average effect size $\pm 1.96 \times \tau$.
*$p < .10$. **$p < .05$. ***$p < .01$.

average effect size for programs that included only professional development or only new curriculum materials.

Features That Moderate Program Impacts

Next, we turn to features that may moderate impacts on student outcomes, beginning with the professional development models. We did not find a significant relationship between professional development contact hours and student outcomes (see Table 5). We first examined whether there was a linear association between contact hours and effect size (column 1). Next, to look for nonlinearities in the contact hours data, the model presented in column 2 compares programs with few hours (i.e., below the 25th percentile; omitted) and those with a larger number of contact hours (i.e., 25th to 50th percentile, 50th to 75th percentile, and 75th percentile and above). We did not find evidence of a significant association in either model. To investigate a threshold effect, the model presented in column 3 examines programs with 16 hours or more and found no relationship; this finding held at various threshold levels (not shown). In a separate analysis, we did not find a significant relationship between the time span (e.g., 1 month, 1 year) over which professional development activities were conducted and effect sizes (models not shown).

TABLE 3

RVE Results Including Controls for Study Design, Study Sample, Subject Area, and Outcome Measure Type

	Dependent variable: Effect size (Hedges's *g*)
Between-study effects	
RCT	−0.022
	(0.104)
State standardized test	−0.264***
	(0.055)
Other standardized test	−0.277***
	(0.053)
Grade—preschool	0.133
	(0.084)
Effect size adjusted for covariates	−0.040
	(0.055)
Subject matter—math	−0.009
	(0.044)
Within-study effects	
State standardized test	−0.208***
	(0.052)
Other standardized test	−0.203***
	(0.059)
Constant	0.395***
	(0.111)
N effect sizes	258
N studies	95
τ^{2} [a]	0.025
Results of joint *F* test[b]	$F = 5.75$, $df = 26.9$, $p < .001$

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. RVE = robust variance estimation; RCT = randomized controlled trial.
[a]τ^2 is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).
[b]Results of the joint *F* test are from a test of the joint significance of all study characteristics included in the model. The *F* test was estimated using the *robumeta* and *clubSandwich* package in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).
*$p < .10$. **$p < .05$. ***$p < .01$.

We next examined the associations between the *focus* of the professional development and effect sizes via a series of multilevel regression models (Table 6). Average effect sizes were

TABLE 4

RVE Results Including Intervention Characteristics (Professional Development and/or Curriculum Materials) as Moderators

	Dependent variable: Effect size (Hedges's *g*)		
Between-study effects			
Professional development only	−0.084	−0.090*	
	(0.051)	(0.052)	
New curriculum materials only		−0.129	
		(0.118)	
Professional development only/new curriculum materials only			−0.099**
			(0.046)
N effect sizes	258	258	258
N studies	95	95	95
τ^{2} [a]	0.025	0.025	0.025
Raw mean effect size			
Professional development only	0.254		
New curriculum materials only	0.091		
Professional development only/new curriculum materials only	0.214		
Both professional development and new curriculum materials	0.252		

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. Models include controls for the following: RCT, state standardized test, other standardized test, grade–preschool, effect size adjusted for covariates, and subject matter. RVE = robust variance estimation; RCT = randomized controlled trial.

[a] τ^{2} is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

*$p < .10$. **$p < .05$. ***$p < .01$.

larger when programs focused on how to use curriculum materials (+0.12 *SD,* $p < .10$) and when they focused on improving teachers' content and pedagogical content knowledge and/or how students learned the content (+0.09 *SD,* $p < .05$). Both of these associations remained significant and similar in size when all predictors were included in the model. Average effect sizes were also larger when the program focused on formative assessment (+0.13 *SD,* $p < .10$), although this did not retain its size or significance in the final model. A focus on content-generic instructional strategies was not a significant predictor of effect size magnitude. Finally, we also found that programs that included more of the foci listed in Table 6 had larger effect sizes, on average, compared with programs focused on a more narrow set of topics (+0.08 *SD,* $p < .01$).

Next, we turned to professional development *activities* (Table 7). No activities for which we coded—observing demonstrations, reviewing generic student work (problems or investigations completed by students outside the teachers' classes), solving math or science problems/working through student materials, developing curriculum or lesson plans, and reviewing teachers' own students' work—were significant predictors of effect size magnitude. However, we find that the number of PD activities incorporated in the program, defined as the total number of the five activities listed in Table 7, significantly predicted effect size magnitude (+0.05 *SD,* $p < .10$).

Table 8 examines the relationship between effect sizes and professional development *formats*. On average, PD programs that had teachers participate alongside other teachers in their school, which we refer to as same-school collaboration, yielded outcomes 0.12 *SD* larger ($p < .10$) than programs without such collaboration. This result is significant in the final model only. In addition, programs that included PD with implementation meetings yielded significantly larger average effect sizes on average than without these meetings (+0.12 *SD* in the final model, $p < .05$).

TABLE 5

RVE Results Including Professional Development Contact Hours as Moderators

	Dependent variable: Effect size (Hedges's *g*)		
Between-study effects			
PD contact hours[a]	0.005		
	(0.007)		
PD contact hours between 25th and 50th percentile (16–34.5 hours)		0.012	
		(0.050)	
PD contact hours between 50th and 75th percentile (35–68 hours)		−0.033	
		(0.065)	
PD contact hours above 75th percentile (>68 hours)		0.069	
		(0.067)	
PD contact hours at or above 25th percentile (>16 hours)			0.017
			(0.043)
N effect sizes	231	231	231
N studies	85	85	85
τ^{2} [b]	0.022	0.024	0.023
Results of joint *F* test[c]	$F = 0.647$, $df = 35.2$, $p = .590$		

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. All models include only studies and/or treatment arms with a professional development component. All models include controls for the following: RCT, state standardized test, other standardized test, grade–preschool, effect size adjusted for covariates, and subject matter. RVE = robust variance estimation; PD = professional development; RCT = randomized controlled trial.

[a]PD contact hours measured as raw PD contact hours/10.

[b]τ^2 is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

[c]Results of the joint *F* test are from a test of the joint significance of the PD contact hours predictors from the second model. The *F* test was estimated using the *robumeta* and *clubSandwich* package in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).

*$p < .10$. **$p < .05$. ***$p < .01$.

These implementation meetings typically allowed teachers to convene briefly with other activity participants to troubleshoot and discuss obstacles and aids to putting the program into practice. Meanwhile, programs that included professional development that incorporated an online component yielded significantly smaller impacts, on average, relative to programs that did not involve any online components (−0.15 *SD* in the final model, $p < .05$). In addition, in the final model, programs that included a summer workshop had larger effect sizes, on average, as compared with those that lacked this component (+0.07 *SD*, $p < .10$). The remaining formats examined—whether the program featured coaching from experts or professional development led by researchers—were not significant predictors of effect size magnitude. The number of PD features was also not a significant predictor of effect size magnitude.

Table 9 shows the associations between features of new curriculum materials and effect sizes. None of the features examined were significantly associated with average effect sizes.

It is important to note that the above moderator analyses provide estimates of the associations between the presence of specific program features and average effect sizes; however, programs without those specific features that positively predict impacts may still positively impact student outcomes on average. Thus, in Table 10, we present the results of these moderator tests summarized in terms of regression-adjusted mean effect sizes. First, we present mean effect sizes based on subgroup analyses without controls for additional program features. These mean effect sizes are based on unconditional meta-regression models estimated using RVE to account for the nesting of effect sizes within studies. Next, we present mean effect sizes based on conditional meta-regression models corresponding to our main moderation analyses with each predictor included separately and controlling for the program features listed in

TABLE 6

RVE Results Including Professional Development Foci as Moderators

	Dependent variable: Effect size (Hedges's *g*)						
Between-study effects							
Generic instructional strategies	−0.014					−0.074	
	(0.076)					(0.072)	
How to use curriculum materials		0.118*				0.119*	
		(0.060)				(0.063)	
Integrate technology			0.185			0.139	
			(0.117)			(0.102)	
Content-specific formative assessment				0.132*		0.105	
				(0.067)		(0.062)	
Improve pedagogical content knowledge/how students learn					0.094**	0.096**	
					(0.043)	(0.043)	
Number of PD features							0.077***
							(0.024)
N effect sizes	237	237	237	237	237	237	237
N studies	89	89	89	89	89	89	89
$\tau^{2\,a}$	0.025	0.025	0.024	0.024	0.026	0.025	0.023
Results of joint *F* test[b]	$F = 3.89$, $df = 20.4$, $p = .012$						

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. All models include only studies and/or treatment arms with a professional development component. All models include controls for the following: RCT, state standardized test, other standardized test, grade–preschool, effect size adjusted for covariates, and subject matter. RVE = robust variance estimation; PD = professional development; RCT = randomized controlled trial.

[a]τ^2 is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

[b]Results of the joint *F* test are from a test of the joint significance of the predictors for all professional development foci included in the model. The *F* test was estimated using the *robumeta* and *clubSandwich* package in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).

*$p < .10$. **$p < .05$. ***$p < .01$.

Table 3. Finally, we present mean effect sizes corresponding to our final moderation analyses with all predictors within each category entered simultaneously. For brevity, we focus on only those features that were statistically significant predictors of program impact in our main models.

The mean effect sizes presented in Table 10 indicate that the overall impacts of programs with and without the moderators of interest were generally positive. Estimated mean effect sizes for STEM professional development and curriculum improvement programs are generally positive even among programs that lacked the intervention features we identified as associated with larger effect sizes. For example, average effect sizes were positive among programs including either professional development or new curriculum materials but not both components ($\overline{g_c} = 0.16$, $\overline{g_{uc}} = 0.14$, $p_{uc} < .01$), professional development programs that had an online component ($\overline{g_{c+}} = 0.10$, $\overline{g_c} = 0.10$, $\overline{g_{uc}} = 0.12$, $p_{uc} < .05$), and professional development that did not contain a focus on improving teachers' content knowledge, pedagogical content knowledge, or knowledge of how students learn ($\overline{g_{c+}} = 0.18$, $\overline{g_c} = 0.18$, $\overline{g_{uc}} = 0.10$, $p_{uc} < .01$). Although impacts were largest on intervenor-developed assessments, estimated mean effects sizes are still positive for impacts on state standardized assessments ($\overline{g_c} = 0.10$, $\overline{g_{uc}} = 0.06$, $p_{uc} < .01$) and other standardized assessments ($\overline{g_c} = 0.09$, $\overline{g_{uc}} = 0.08$, $p_{uc} < .01$). Furthermore, the differences in mean effect sizes within each category based on estimating unconditional models, with the exception of same-school collaboration, are generally comparable in direction and magnitude with those based on conditional models.

TABLE 7

RVE Results Including Professional Development Activities as Moderators

	Dependent variable: Effect size (Hedges's *g*)						
Between-study effects							
Observed demonstration	0.033					0.025	
	(0.049)					(0.046)	
Reviewed generic student work		0.047				−0.014	
		(0.087)				(0.085)	
Solved problems/worked through student materials			0.077			0.070	
			(0.047)			(0.047)	
Developed curriculum materials/lesson plans				0.064		0.036	
				(0.061)		(0.063)	
Reviewed own student work					0.033	0.017	
					(0.064)	(0.071)	
Number of PD features							0.046*
							(0.024)
N effect sizes	237	237	237	237	236	236	237
N studies	89	89	89	89	88	88	89
τ^{2} [a]	0.025	0.021	0.024	0.025	0.018	0.020	0.021
Results of joint *F* test[b]	$F = 0.883$, $df = 21.2$, $p = .509$						

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. All models include only studies and/or treatment arms with a professional development component. All models include controls for the following: RCT, state standardized test, other standardized test, grade–preschool, effect size adjusted for covariates, and subject matter. RVE = robust variance estimation; PD = professional development; RCT = randomized controlled trial.

[a] τ^{2} is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

[b] Results of the joint *F* test are from a test of the joint significance of the predictors for all professional development activities included in the model. The *F* test was estimated using the *robumeta* and *clubSandwich* package in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).

*$p < .10$. **$p < .05$. ***$p < .01$.

Study Design Moderators

Next, we examined whether a wider range of study design characteristics and implementation contexts were associated with effect size magnitudes (Table 11). Studies that included credit incentives for participating teachers showed smaller impacts on average (−0.18 *SD* in the final model, $p < .05$). We found that the unit of assignment to the program (schools vs. teachers) did not predict effect sizes. We did not find significant relationships between effect size magnitudes and whether teacher participation was voluntary or mandatory, or whether teachers received monetary incentives; however, this information was unreported in many studies. Finally, we found that the level of attrition did not predict the size of impacts.

To further examine the potential role of attrition bias, we additionally examined the adjusted mean effect sizes for studies that do and do not meet our attrition standards. We see little evidence that studies with attrition problems reported larger impacts (see Supplemental Appendix Table A2 in the online version of the journal).

In separate analyses, no other study design features were significantly related to outcomes, including whether the study involved one district, multiple districts, and/or states; whether the study was conducted in the United States or abroad; whether the study was conducted in an urban versus nonurban setting; and whether the study sample was majority low income. We also found that study size (defined as the average number of treatment clusters across effect sizes within a study) was not a significant predictor of study outcomes (results not shown).

TABLE 8

RVE Results Including Professional Development Formats as Moderators

	Dependent variable: Effect size (Hedges's *g*)							
Between-study effects								
Same-school collaboration	0.109						0.123*	
	(0.076)						(0.067)	
Implementation meetings		0.085*					0.117**	
		(0.049)					(0.045)	
Any online PD			−0.161***				−0.153**	
			(0.053)				(0.057)	
Summer workshop				0.093**			0.074*	
				(0.043)			(0.038)	
Expert coaching					0.035		0.053	
					(0.049)		(0.051)	
PD leaders—researchers						−0.019	−0.037	
						(0.043)	(0.038)	
Number of PD features								0.023
								(0.024)
N effect sizes	237	237	237	236	237	231	231	237
N studies	89	89	89	88	89	86	86	89
τ^2 [a]	0.020	0.022	0.024	0.023	0.026	0.025	0.015	0.023
Results of joint *F* test[b]	$F = 2.72$, $df = 26.2$, $p = .035$							

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. All models include only studies and/or treatment arms with a professional development component. All models include controls for the following: RCT, state standardized test, other standardized test, grade–preschool, effect size adjusted for covariates, and subject matter. RVE = robust variance estimation; PD = professional development; RCT = randomized controlled trial.

[a] τ^2 is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

[b] Results of the joint *F* test are from a test of the joint significance of the predictors for all professional development formats included in the model. The *F* test was estimated using the *robumeta* and *clubSandwich* package in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).

*$p < .10$. **$p < .05$. ***$p < .01$.

Publication Bias

Finally, we considered whether effect sizes from peer-reviewed sources differed in magnitude, on average, relative to effect sizes from other sources. In Table 12, we present results from a meta-regression model estimated using RVE that includes whether the effect size was from a peer-reviewed publication as a moderator. We found that effect sizes from peer-reviewed studies were larger by 0.05 (uncontrolled, Model 1) or 0.07 *SD* (controlled, Model 2) than those from other studies when using the RVE approach and after controlling for study design, study sample, subject area, and outcome measure type. However, these differences were not statistically significant.

We also conducted two additional tests for publication bias. First, we assessed publication bias using Egger's regression test (Egger, Smith, Schneider, & Minder, 1997). Given the multiple effect sizes within each study, we conducted this test at the study level by regressing the study average standard normal deviation (the average effect size divided by the average standard error) on the inverse of the study average effect size standard error. This approach tests the null hypothesis that the intercept is zero; if the null hypothesis is rejected, this indicates that smaller (or less precise) studies have systematically larger or smaller reported effect sizes relative to larger (or more precise) studies, which could be due to publication bias. We then use the "trim and fill" method to examine the magnitude of potential publication bias (Duval & Tweedie, 2000). Although the models estimated are not precisely analogous to our preferred estimation

TABLE 9

RVE Results Including Characteristics of Interventions Involving New Curriculum Materials as Moderators

	Dependent variable: Effect size (Hedges's *g*)				
Between-study effects					
Implementation guidance	0.060				0.062
	(0.053)				(0.056)
Laboratory/hands-on experience, curriculum kits		−0.029			−0.026
		(0.055)			(0.054)
Curriculum dosage (number of hours)			0.000		0.000
			(0.001)		(0.001)
Curriculum proportion replaced (0.00–1.00)				−0.060	
				(0.151)	
N effect sizes	193	193	193	192	193
N studies	77	77	77	76	77
τ^{2} [a]	0.027	0.031	0.030	0.025	0.031
Results of joint *F* test[b]	$F = 0.415$, $df = 24.4$, $p = .744$				

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. All models include only studies and/or treatment arms that included new curriculum materials. All models include controls for the following: RCT, state standardized test, other standardized test, grade–preschool, effect size adjusted for covariates, and subject matter. RVE = robust variance estimation; RCT = randomized controlled trial.

[a]τ^2 is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

[b]Results of the joint *F* test are from a test of the joint significance of the predictors for all characteristics of interventions involving new curriculum materials included in the model. The *F* test was estimated using the *robumeta* and *clubSandwich* package in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).

*$p < .10$. **$p < .05$. ***$p < .01$.

approach, results indicate potential publication bias in the full sample as demonstrated by the fact that studies with larger effect size standard errors (i.e., smaller and less precise studies) have larger effects ($p < .01$). A comparison of the adjusted and unadjusted estimated average effect sizes from the trim-and-fill method indicates that the magnitude of potential publication bias is substantial. Second, to account for the nested structure of the data, we also used a modification of the Egger's regression test by adding the standard errors of the effect sizes as a moderator to the unconditional RVE meta-regression model. If effect size standard errors are significant predictors of effect sizes, this could similarly indicate the presence of publication bias. As above, results indicate potential publication bias in the full sample ($p < .01$).

However, results of conducting both methods separately for peer-reviewed and non-peer-reviewed studies detect potential publication bias only among peer-reviewed studies. Results of both Egger's regression test and the RVE approach indicate that smaller (or less precise) studies report larger impacts ($p < .01$). However, there is less evidence of systematic differences in reported effects sizes based on study size or precision among studies from other sources ($p = .12$, $p = .05$). This suggests that the association between study size or precision and reported impacts among peer-reviewed studies may be a result of publication bias rather than other factors (e.g., more effective interventions are more costly and, therefore, evaluated with small samples). This highlights the importance of the inclusion of the "grey literature" in systematic reviews and meta-analyses of instructional improvement efforts (Polanin, Tanner-Smith, & Hennessy, 2016). For full results of these tests, see Supplemental Appendix Table A3 and Figure A1 in the online version of the journal.

Sensitivity Checks

We conducted additional analyses to examine the robustness of our results. First, we address the fact that in 20 cases, authors reported some

TABLE 10

Regression-Adjusted Mean Effect Sizes Based on Unconditional and Conditional RVE Meta-Regression Models

	Subgroup analysis (unconditional RVE model)	Conditional RVE model		Conditional RVE model with controls for other program features	
	$\overline{g_{uc}}$[a], p_{uc}[b]	$\overline{g_c}$[c]	p_c[d]	$\overline{g_{c+}}$[e]	p_{c+}[f]
Overall effect size	0.209***	—	—	—	—
Program type					
Professional development and curriculum materials	0.235***	0.254	**	—	—
Professional development only/curriculum materials only	0.136***	0.156		—	—
Outcome type					
State standardized test	0.060***	0.101	***	—	—
Other standardized test	0.084***	0.087	***	—	—
Intervenor-developed test	0.371***	0.365		—	—
Program feature category: Professional development focus					
Focus on how to use curriculum materials					
Yes	0.228***	0.258	*	0.260	*
No	0.134***	0.140		0.141	
Focus on content-specific formative assessment					
Yes	0.387***	0.340	*	0.321	
No	0.178***	0.208		0.216	
Focus on improving content knowledge/pedagogical content knowledge/how students learn					
Yes	0.303***	0.269	**	0.275	**
No	0.095***	0.175		0.179	
Program feature category: Professional development formats					
PD includes same-school collaboration					
Yes	0.198***	0.248		0.247	*
No	0.263***	0.139		0.125	
PD includes implementation meetings					
Yes	0.283***	0.284	*	0.297	**
No	0.169***	0.199		0.180	
Any online professional development					
Yes	0.116**	0.103	***	0.096	**
No	0.235***	0.264		0.249	
PD includes summer workshop					
Yes	0.218***	0.266	**	0.253	*
No	0.190***	0.174		0.179	

Note. First column: Regression-adjusted mean effect sizes are based on subgroup analyses. Unconditional RVE meta-regression models were estimated including only effect sizes with the row feature. Second and third columns: Regression-adjusted mean effect sizes are based on the results of estimating conditional RVE meta-regression models. Models included each of the row feature separately as a moderator. All models included controls for the variables listed in Table 3. Regression-adjusted mean effect sizes were calculated using the overall average of the study-level values of each included covariate. Fourth and fifth columns: Regression-adjusted mean effect sizes are based on the results of estimating conditional RVE meta-regression models. Models included all row features in a given category simultaneously as moderators. All models included controls for the variables listed in Table 3. Regression-adjusted mean effect sizes were calculated using the overall average of the study-level values of each included covariate. RVE = robust variance estimation; PD = professional development.

[a] $\overline{g_{uc}}$ = estimated mean effect size from unconditional RVE meta-regression model.

(continued)

TABLE 10 (CONTINUED)

[b] p_{uc} = p value on constant from unconditional meta-regression model.

[c] $\overline{g_c}$ = estimated regression-adjusted mean effect size from conditional RVE meta-regression model.

[d] p_c = p value on coefficient for program feature on interest from meta-regression model.

[e] $\overline{g_{c+}}$ = estimated regression-adjusted mean effect size from conditional RVE meta-regression model including all row features in a given category simultaneously as moderators.

[f] p_c+ = p value on coefficient for program feature on interest from meta-regression model including all row features in a given category simultaneously as moderators.

*$p < .10$. **$p < .05$. ***$p < .01$.

TABLE 11

RVE Results Including Other Research Design Elements as Moderators

	Dependent variable: Effect size (Hedges's *g*)							
Between-study effects								
Unit of assignment is teacher	−0.054							−0.060
	(0.043)							(0.052)
Business as usual control group		0.008						0.076
		(0.079)						(0.085)
Teacher participation—voluntary			−0.035					−0.032
			(0.045)					(0.054)
Teacher participation—missing			0.111					0.114
			(0.089)					(0.103)
Teacher incentive—credit				−0.167**				−0.177*
				(0.074)				(0.097)
Teacher credit incentive—missing				−0.050				−0.137
				(0.055)				(0.097)
Teacher incentive—monetary					−0.054			0.016
					(0.042)			(0.057)
Teacher monetary incentive—missing					−0.019			0.073
					(0.063)			(0.110)
High cluster attrition						−0.056		0.022
						(0.051)		(0.057)
Cluster attrition—missing						−0.070		−0.069
						(0.064)		(0.078)
High student attrition							−0.068	−0.054
							(0.046)	(0.052)
Student attrition—missing							−0.031	0.031
							(0.054)	(0.065)
N effect sizes	258	258	258	258	258	258	258	258
N studies	95	95	95	95	95	95	95	95
τ^2 [a]	0.024	0.022	0.026	0.024	0.026	0.026	0.027	0.028
Results of joint *F* test[b]	$F = 1.30$, $df = 23.2$, $p = .284$							

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. All models include controls for the following: RCT, state standardized test, other standardized test, grade–preschool, effect size adjusted for covariates, and subject matter. RVE = robust variance estimation; RCT = randomized controlled trial.

[a] τ^2 is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

[b] Results of the joint *F* test are from a test of the joint significance of the predictors for all additional research design elements in the model. The *F* test was estimated using the *robumeta* and *clubSandwich* package in R (Fisher & Tipton, 2015; Tipton & Pustejovsky, 2015).

*$p < .10$. **$p < .05$. ***$p < .01$.

TABLE 12

RVE Results Including Peer-Reviewed Publication Type as a Moderator

	Dependent variable: Effect size (Hedges's *g*)	
Between-study effects		
Peer-reviewed source	0.046	0.067
	(0.050)	(0.041)
N effect sizes	258	258
N studies	95	95
τ^{2} [a]	0.038	0.026

Note. We assume that the average correlation between all pairs of effect sizes within studies is .80. Models in column 2 include controls for the following: RCT, state standardized test, other standardized test, grade–preschool, effect size adjusted for covariates, and subject matter. RVE = robust variance estimation.

[a] τ^2 is the method-of-moments estimate of the between-study variance in the underlying effects provided by the *robumeta* package in Stata 15 (Tanner-Smith & Tipton, 2014).

*$p < .10$. **$p < .05$. ***$p < .01$.

impacts as "not statistically significant" and did not provide enough information to calculate an effect size. A concern is that failing to provide this information may be correlated with program features. In addition, those effect sizes that are not statistically significant but negative in sign may be less frequently reported than effect sizes that are positive in sign. We, therefore, tested the robustness of our results to the inclusion of these 20 missing effect sizes by imputing a range of values for missing effect sizes ($g = 0.00$, $g = -0.10$, and $g = -0.20$) and using the study-level mean of the effect size standard error to calculate their weights. If our results are driven by differential reporting of information regarding effect sizes that are not statistically significant based on whether point estimates are positive or negative in sign, including these impacts should attenuate our results. In general, including these effect sizes did not substantively change our results. In the majority of cases where our primary results showed a significant association between program features and outcomes, results are comparable in magnitude and remain statistically significant. The only exceptions are the number of professional development activities and use of same-school collaboration, which are no longer significant predictors of program impact, although the associations are comparable in magnitude with our main estimates.

We also found that our results were largely robust to excluding studies in settings outside the United States and to excluding studies with weaker designs (i.e., studies that were not RCTs). Associations are generally comparable in magnitude with our main results, although in some cases less precisely estimated. The only exceptions are that a focus on how to use new curriculum materials is not a significant predictor of impacts after excluding non-RCT studies, and a focus on how to use new curriculum materials and the number of professional development activities are not significant predictors of impacts after excluding studies outside the United States. These associations are comparable in magnitude with our main estimates. We also examined the sensitivity of our results to choosing different values of the within-study correlation between effect sizes. In our primary models, we specify the correlation to be .80, which is the value recommended by Tanner-Smith and Tipton (2014). Using alternative specifications of $\rho = 0.50$, 0.70, and 0.90 did not change our results. Full results of these sensitivity checks are available from the authors on request.

Discussion and Conclusion

To summarize, we found that studies of STEM instructional improvement programs had, on average, positive effects on student achievement, with a mean pooled effect size across studies of 0.21 *SD*. Compared with those found in prior reviews, these pooled effect sizes are in the middle range, smaller than those identified in the Yoon et al. review (0.57 *SD* in math and 0.51 *SD* in science) and the Taylor et al. review (0.49 *SD* in science), and somewhat larger than those identified in the Scher and O'Reilly review (0.14 *SD* in math and 0.13 *SD* in science). Although the effect sizes differ, our results confirm earlier reviewers' findings that studies of STEM instructional improvement efforts tend to show positive results.

We conducted a series of analyses to examine the relationships between program characteristics and the size of achievement impacts. The characteristics that were significantly associated

with improved student learning across the current set of studies included the following:

- the use of professional development along with new curriculum materials;
- a focus on improving teachers' content/pedagogical content knowledge, or understanding of how students learn; and
- specific formats, including
  - meetings to troubleshoot and discuss classroom implementation of the program,
  - the provision of summer workshops to begin the professional development learning process, and
  - same-school collaboration.

We also found that, on average, programs that provided any component of the PD online had poorer student outcomes than programs that did not use an online PD component. In general, there was not a statistically significant difference in the magnitudes of these associations depending on whether programs focused on mathematics or science.

Components Associated With STEM Program Effectiveness

Taken together, we generally find that providing teachers with opportunities to learn about the materials they will use with students and/or to participate in programs that seek to improve their content or pedagogical content knowledge is associated with improved student outcomes. These findings accord with prior cross-sectional research. For example, Boyd, Grossman, Lankford, Loeb, and Wyckoff (2009) found that teacher preparation programs that focused closely on teachers' classroom practice, including reviewing the district curriculum, produced teachers with better student outcomes in their first year of teaching. Cohen and Hill (1998) also found that teachers' participation in curriculum-centered professional development was related to student achievement. These findings also lend support to the conclusions drawn in prior reviews conducted by Scher and O'Reilly (2009) and Kennedy (1999). Scher and O'Reilly (2009) found that interventions that focused on both content and pedagogy posted stronger student outcomes as compared with interventions that focused on pedagogy alone, whereas Kennedy (1999) concluded that programs were more effective when they focused on how to teach specific content and on how students learn the same content.

Most studies of curriculum materials in our data set included at least some component of professional development, and vice versa. However, examining studies that included both elements jointly leads us to see that, on average, programs that incorporated both professional development and new curriculum materials had larger impacts as compared with programs that included only one of these components. These findings lend support to the argument advanced by some scholars that curriculum materials alone, even those designed to be educative and supportive of teacher implementation, may be insufficient to change teaching practice, given the complexity of classroom instructional interactions, and the often-ingrained nature of traditional inquiry–response–evaluation teaching practices (Alozie, Moje, & Krajcik, 2010). Meanwhile, perhaps when professional development is provided without reference to specific curriculum, teachers may struggle to implement what they have learned while using existing curricular materials and textbooks.

We also found a positive association between student outcomes and teachers' participation in implementation meetings, which were defined as meetings in which teachers met formally or informally with other activity participants to discuss enacting intended practices. We also found a significant relationship, in our final model, between same-school collaboration (teachers participating in PD alongside their colleagues) and student outcomes. These findings align with prior work (e.g., Darling-Hammond & McLaughlin, 1995), emphasizing the importance of providing teachers with opportunities to discuss instructional innovations with colleagues (e.g., Penuel, Sun, Frank, & Gallagher, 2012) and discuss and troubleshoot issues that arise when implementing new instructional approaches.

Perhaps more surprisingly, the inclusion of a summer workshop was positively related to student outcomes. Prior syntheses (Scher & O'Reilly, 2009; Yoon et al., 2007) did not

examine this variable specifically. The summer workshop format for professional development has been critiqued in the past for its typically "one-shot" nature, but it may be the case that an intensive summer professional development provides an effective "springboard" for school-year implementation. However, programs in which participants completed a portion of the professional development online had weaker student outcomes, on average, as compared with programs that did not include any online PD. Dede, Ketelhut, Whitehouse, Breit, and McCloskey (2009) point out that although online teacher professional development is burgeoning, relatively little rigorous research has examined its effectiveness. Although these analyses are correlational in nature, the positive results associated with both the summer professional development and the teacher implementation meetings are consistent with the notion that teachers may have benefited from both intensive and ongoing opportunities to interact with one another around program content, which may have been less salient in the online format.

In contrast to two earlier reviews (Scher & O'Reilly, 2009; Yoon et al., 2007), we find no evidence of a positive association between the duration of professional development and program impacts. Yoon compared the effectiveness of interventions that included greater than with less than 14 hours of professional development, finding that the former had larger impacts on student achievement. Scher and O'Reilly (2009) compared the effectiveness of interventions conducted over two or more years with those conducted over 1 year and found that those conducted over two or more years were more effective, on average, among math-focused, but not science-focused, interventions. However, both these reviews note that the small number of included studies limited their ability to draw firm conclusions. The current findings, using a continuous measure of contact hours and a separate measure of time span, suggest that programs that were limited in duration, nonetheless, generally had positive impacts on average. For example, several programs that combined new curriculum materials with a short amount of professional development documented moderate to large impacts on student achievement (e.g., Arnold, Fisher, Doctoroff, & Dobbs, 2002; Clements & Sarama, 2007; Presser, Vahey, & Dominguez, 2015). In contrast, some studies of highly intensive professional development programs showed little or no impacts on student learning (e.g., Devlin-Scherer et al., 1998; Jacob, Hill, & Corey, 2017; Van Egeren et al., 2014). Our findings echo those of Kennedy's (1999, 2016), who did not find a clear benefit of contact hours or program duration, and concluded that the core condition for program effectiveness was valuable content; more hours of a given intervention will not help if the intervention content is not useful.

Similar to Taylor et al. (2018), our analysis did not detect any relationship between student outcomes and study design, science subject matter, and grade level. However, we found that average impacts varied considerably depending on the type of student test used. On average, student outcomes were larger on researcher-designed assessments as compared with standardized tests. Although standardized tests may have benefits such as face validity and broad content representation, it may be that standardized tests are not especially sensitive to instructional improvement efforts, due to differences in the skills measured by the tests versus those targeted in the intervention (see also C. J. Hill et al., 2008; Sussman & Wilson, 2019; Taylor et al., 2018). If this is the case, to understand how instructional improvement interventions influence student learning, researchers may need to include both assessments of outcomes that are closely tied to the student learning goals along with broader standardized tests.

Limitations and Future Directions

We note several limitations to the current review. One limitation is missing data. First, although we attempted to search both the published and unpublished literatures, studies may have eluded our grasp. Second, many programs and interventions that are routinely conducted in schools are never evaluated, and these programs may differ in unknown ways from those that are formally evaluated. Most of the programs studied are boutique programs, often designed, operated, and evaluated by university or contract

researchers. We know little about the efficacy of professional development that reaches typical teachers. The characteristics we identified as potentially effective here may not carry over to typical conditions and with the resources conventionally available in school districts.

Missing data in study reports posed another limitation. We initially hoped to code the included studies for a number of additional features that have been hypothesized in the literature to influence the effectiveness of instructional improvement programs, including district and school leadership support, competing instructional improvement initiatives, teacher recruitment methods, and the financial resources provided to support the intervention (S. M. Wilson, 2013). However, we found that few study reports contained sufficient detail on these matters, making tracking their impact impossible. This omission is striking; in nearly all published and unpublished reports, the district context is simply a black box, or merely a site for teacher recruitment and service delivery. The need for the inclusion of more contextual information in study reports is pressing, especially because many district administrators evaluate proposed programs based on the extent to which studies were carried out in "districts like ours." Gathering contextual information would also provide insight into the conditions necessary for instructional improvement programs to thrive.

In addition, it is important to note that, as is generally the case in meta-analyses, the moderator analyses we have conducted are correlational. The authors of the included studies generally did not randomly manipulate the variables that we identified as associated with improved student outcomes in the moderator analyses, such as by randomly assigning teachers to participate versus not participate in a summer workshop. As a result, moderator analyses are potentially confounded by unobserved or inadequately measured study and student characteristics. The characteristics of professional development and curriculum programs that appear promising in the current study's moderator analyses, thus, are not definitive, but point toward potentially productive areas for future experimental work. Future randomized experiments comparing instructional programs with and without these components are warranted to estimate causally the effects of robust STEM professional development and curriculum interventions on student learning.

Despite these limitations, however, we were able to distill findings from dozens of recent experimental and strong quasi-experimental evaluations of instructional improvement programs in STEM. Based on the extant literature, the types of practices that were associated in our review with improved student learning, such as providing teachers with opportunities to engage with the curriculum they teach, develop their content knowledge, pedagogical content knowledge, and understanding of how students learn, and discuss classroom implementation, likely occur infrequently in the instructional improvement programs typically offered in U.S. schools (e.g., H. C. Hill, 2009). Future experimental studies that build on the current findings are needed to advance researchers' and policymakers' understanding of core instructional reform practices that improve student learning in STEM.

Acknowledgments

The authors are grateful to Betsy Becker, Hilda Borko, Joan Heller, James S. Kim, and Jeremy Roschelle for helpful comments. Thanks to Patrick Aquino, Wilson Boardman, Jonathan Hamel Sellman, Andrea Humez, Meghan Kelly, Jiwon Lee, Angie Luo, Manish Parmar, and Melanie Rucinski for their assistance with data preparation and analysis.

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: This material is based upon work supported by the National Science Foundation under Grant #1348669. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

ORCID iD

H. C. Hill https://orcid.org/0000-0001-5181-1573

Notes

1. For example, in the area of middle and secondary math curriculum materials, Slavin and colleagues searched the published and gray literature dating back to 1970, conducting

> [a] broad literature search . . . in an attempt to locate every study that could possibly meet the inclusion requirements. This included obtaining all of the middle school studies cited by the What Works Clearinghouse (2008b) and the middle and high school studies cited by NRC (2004), by Clewell et al. (2004), and by other reviews of mathematics programs, including technology programs that teach math (e.g., Chambers, 2003; Kulik, 2003; Murphy et al., 2002). Electronic searches were made of educational databases (JSTOR, Education Resources Information Center, EBSCO, PsycINFO, Dissertation Abstracts), Web-based repositories (Google, Yahoo, Google Scholar), and education publishers' Web sites. Citations of studies appearing in the first wave of studies were also followed up. A particular effort was made to find non-U.S. studies.They performed a similar search for elementary math curriculum materials (Slavin & Lake, 2008), elementary science curriculum materials (Slavin, Lake, Hanley, & Thurston, 2014), and secondary science materials (Cheung et al., 2017). We refer the reader to those articles for specifics. In the area of professional development, Yoon et al. (2007) searched for studies dating from 1986 and explain that

> Studies were gathered through an extensive electronic search of published and unpublished research literature. The review protocol included a list of keywords that guided the literature search. Seven electronic databases were core data sources: ERIC, PsycINFO, ProQuest, EBSCO's Professional Development Collection, Dissertation Abstracts, Sociological Collection, and Campbell Collaboration. These databases were searched separately for each of the three subjects under review (mathematics, science, and reading and English/language arts). In consultation with a reference librarian, search parameters were developed using database-specific keywords . . . A deliberately wide net captured literature on professional development and student achievement, broadly defined . . . Fourteen key researchers were also asked to identify research for the study. Eight researchers responded, recommending additional studies that fit the study purpose. Finally, existing literature reviews and research syntheses were consulted to ensure that no key studies were omitted.

2. As another point of comparison, the What Works Clearinghouse protocols have generally limited their scope to studies from the past 20 years, with some categories such as conference proceedings limited to the past 7 years (Brown, Card, Dickersin, Greenhouse, Kling, & Littell, 2008).

3. The specific search strings applied to searches of the titles and abstracts were as follows: ("professional development" OR "faculty development" OR "Staff development" OR "teacher improvement" OR "inservice teacher education" OR "peer coaching" OR "teachers' institute*" OR "teacher mentoring" OR "Beginning teacher induction" OR "teachers' Seminar*" OR "teachers' workshop*" OR "teacher workshop*" OR "teacher center*" OR "teacher mentoring" OR curriculum OR instruction*) AND ("Student achievement" OR "academic achievement" OR "mathematics achievement" OR "math achievement" OR "science achievement" OR "Student development" OR "individual development" OR "student learning" OR "intellectual development" OR "cognitive development" OR "cognitive learning" OR "Student Outcomes" OR "Outcomes of education" OR "educational assessment" OR "educational measurement" OR "educational tests and measurements" OR "educational indicators" OR "educational accountability") AND ("*experiment*" OR "control*" OR "regression discontinuity" OR "compared" OR "comparison" OR "field trial*" OR "effect size*" OR "evaluation") AND ("Math*" OR "*Algebra*" OR "Number concepts" OR "Arithmetic" OR "Computation" OR "Data analysis" OR "Data processing" OR "Functions" OR "Calculus" OR "Geometry" OR "Graphing" OR "graphical displays" OR "graphic methods" OR "Science*" OR "Data Interpretation" OR "Laboratory Experiments" OR "Laboratory Procedures" OR "Experiment*" OR "Inquiry" OR "Questioning" OR "investigation*" OR "evaluation methods" OR "laboratories" OR "biology" OR "observation" OR "physics" OR "chemistry" OR "scientific literacy" OR "scientific knowledge" OR "empirical methods" OR "reasoning" OR "hypothesis testing").

4. As Slavin (2008) defined this contrast,

> A program is defined here as any set of replicable procedures, materials, professional development, or service configurations that educators could choose to implement to improve student outcomes. A program is distinct from a variable in consisting of a specific, well-specified set of procedures and supports. Class size, assigning homework, or provision of bilingual education are variables, for example, whereas programs typically are based on particular textbooks, computer software, and/or instructional processes and usually have a name and a specific provider, such as a company, university, or individual.

5. For studies with multiple treatment arms, this includes separate effect sizes from each treatment contrast. For studies that reported impacts for multiple groups of teachers or students (e.g., studies that reported impacts separately by grade), this includes separate impacts for each teacher and student sample.

However, we do not include multiple effect sizes for impacts on the same assessment for the same teacher and student sample (e.g., cases where studies administered the same assessment to examine follow-up impacts over time). We also include a separate effect size for each assessment of math or science achievement reported by each study. For example, if a study reported impacts on two separate assessments (e.g., a standardized test and a researcher-designed assessment), both effect sizes were coded and included in the analysis. We did not include impacts on assessment subscales or subscores if impacts on total scores on the assessment were also reported. We included impacts on subscales or subscores only if impacts on total scores were not reported.

6. Our technical advisory board consisted of five members—three with expertise in instructional improvement and program evaluation, and two with expertise in meta-analysis. We met in person with the first group to gather feedback on our proposed coding system. We consulted with and sent drafts of our article to the latter group for feedback on our data and models.

7. Some author-supplied effect sizes were described as only as being similar to a standardized mean difference (including standardized coefficients from multilevel models). These were treated as Cohen's *d* effect sizes and converted to Hedges's *g*.

8. For these cases, a standardized effect size was calculated by the ratio of the unstandardized regression coefficient and an estimate of the standard deviation of the outcome. The outcome standard deviation was estimated using the formula provided by Higgins and Deeks (2008):

$$SD = \frac{SE}{\sqrt{\frac{1}{N_T} + \frac{1}{N_c}}},$$

where N_T represents the number of students in the treatment group and N_C represents the number of students in the control group, and *SE* represents the coefficient standard error.

9. These include cases where the authors did not take into account the nesting of students within classrooms and/or schools, cases where effect sizes were reported without the associated standard errors, and cases where the authors reported cluster-adjusted impact estimates, but it was necessary to calculate standardized effect sizes based on other information. For example, this includes cases where the authors reported that cluster-adjusted regression results were significant below a given value (e.g., $p < .05$) but standard errors, *p* values, or other test statistics (e.g., *t* statistics) were not reported. Of the 258 effect sizes included in this study, we applied a correction to adjust the standard error for clustering in 114 cases. The majority of effect sizes that did not require the standard error correction for clustering were based on results of multilevel models and regression models with clustered standard errors (e.g., using *t* statistics, standard errors and regression coefficients, and *p* values). Other effect sizes were reported at the cluster level (e.g., differences in mean classroom performance); no clustering adjustment was necessary in these cases.

10. These include 20 cases where the authors referred to "no significant effect" on one or more outcomes but did not report an effect size, and nine cases where authors reported some outcome information but did not provide enough information to calculate an effect size (e.g., cases where the outcome information included only raw means).

11. Specifically, we assumed that the standardized effect size (Hedges's *g*) took on a range of values ($g = 0.00$, $g = -0.10$, and $g = -0.20$) and used the study-level mean standard error based on nonmissing effect sizes.

12. Some studies contained multiple treatment arms where the characteristic was present in at least one arm, but not present in at least one other arm. In Table 1, we consider whether the feature was present in any treatment arm of the study. In all subsequent analyses, we use the study-level mean.

References

Agodini, R., Harris, B., Remillard, J., & Thomas, M. (2013). *After two years, three elementary math curricula outperform a fourth*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Alozie, N. M., Moje, E. B., & Krajcik, J. S. (2010). An analysis of the supports and constraints for scientific discussion in high school project-based science. *Science Education*, *94*, 395–427.

Arnold, D. H., Fisher, P. H., Doctoroff, G. L., & Dobbs, J. (2002). Accelerating math development in Head Start classrooms. *Journal of Educational Psychology*, *94*, 762–770.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—The role of curriculum materials in teacher learning and instructional reform? *Educational Researcher*, *25*(9), 6–14.

Banilower, E., Smith, P. S., Weiss, I. R., & Pasley, J. D. (2006). The status of K-12 science teaching in the United States: Results from a national observation survey. In D. Sunal & E. Wright (Eds.), *The impact of the state and national standards on K-12 science teaching* (pp. 83–122). Greenwich, CT: Information Age.

Becker, K., & Park, K. (2011). Effects of integrative approaches among science, technology, engineer-

ing, and mathematics (STEM) subjects on students' learning: A preliminary meta-analysis. *Journal of STEM Education: Innovations and Research*, *12*(5/6), 23–37.

Blank, R. K., & de las Alas, N. (2010, March). *Effects of teacher professional development on gains in student achievement: How meta-analysis provides scientific evidence useful to education leaders*. Paper presented at the Society for Research on Educational Effectiveness Spring Conference, Washington, DC.

Borman, K. M., Cotner, B. A., Lee, R. S., Boydston, T. L., & Lanehart, R. (2009, March). *Improving elementary science instruction and student achievement: The impact of a professional development program*. Paper presented at the Society for Research on Educational Effectiveness Spring Conference, Washington, DC.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*, 416–440.

Brown, H., Card, D., Dickersin, K., Greenhouse, J., Kling, J., & Littell, J. (2008). *Report of the What Works Clearinghouse expert panel*. Washington, DC: National Board for Education Sciences.

Cavanagh, S. (2015, August 17). Spending on instructional materials, construction climbing in K-12. *EdWeek*. Retrieved from https://marketbrief.edweek.org/marketplace-k-12/instructional_materials_construction_spending_climbing_in_k-12/

Cervetti, G. N., Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. G. (2012). The impact of an integrated approach to science and literacy in elementary school classrooms. *Journal of Research in Science Teaching*, *49*, 631–658.

Chambers, E. A. (2003). *Efficacy of educational technology in elementary and secondary classrooms: A meta-analysis of the research literature from 1992–2002* (Unpublished doctoral dissertation). Southern Illinois University Carbondale.

Cheung, A., Slavin, R. E., Kim, E., & Lake, C. (2017). Effective secondary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*, *54*, 58–81.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, *86*, 79–122.

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*, 136–163.

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

Cohen, D. K., & Hill, H. C. (1998). *Instructional policy and classroom performance: The mathematics reform in California*. Philadelphia, PA: Consortium for Policy Research in Education.

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

Confrey, J. (2006). Comparing and contrasting the National Research Council report on evaluating curricular effectiveness with the What Works Clearinghouse approach. *Educational Evaluation and Policy Analysis*, *28*, 195–213.

Confrey, J., & Stohl, V. (Eds.). (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: National Academies Press.

Corcoran, T. B. (1995). *Helping teachers teach well: Transforming professional development*. Philadelphia, PA: CPRE Policy Briefs. Retrieved from https://files.eric.ed.gov/fulltext/ED388619.pdf

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

Dede, C., Ketelhut, D., Whitehouse, P., Breit, L., & McCloskey, E. M. (2009). A research agenda for online teacher professional development. *Journal of Teacher Education*, *60*, 8–19.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, 181–199. doi:10.3102/0013189X08331140

Desimone, L. M. (2011). A primer on effective professional development. *Phi Delta Kappan*, *92*(6), 68–71.

Devlin-Scherer, W., Spinelli, A. M., Giammatteo, D., Johnson, C., Mayo-Molina, S., McGinley, P., & Zisk, L. (1998, February). *Action research in professional development schools: Effects on student learning*. Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, New Orleans, LA.

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A. M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, *87*, 243–282.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and

adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

Fisher, Z., & Tipton, E. (2015). *robumeta: An R-package for robust variance estimation in meta-analysis*. Retrieved from https://arxiv.org/pdf/1503.02220.pdf

Frechtling, J., Sharp, L., Carey, N., & Vaden-Kiernan, N. (1995). *Professional development programs: A perspective on the last four decades*. Washington, DC: National Science Foundation, Division of Research, Evaluation and Communication.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82*, 300–329.

Gardella, J. H., Fisher, B. W., & Teurbe-Tolon, A. R. (2017). A systematic review and meta-analysis of cyber-victimization and educational outcomes for adolescents. *Review of Educational Research*, *87*, 283–308.

Garet, M. S., Wayne, A. J., 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.

Gersten, R., Taylor, M. J., Keys, T. D., Rolfhus, E., & Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches* (REL 2014–010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs

Goldenberg, C., & Gallimore, R. (1991). Changing teaching takes more than a one-shot workshop. *Educational Leadership*, *49*(3), 69–72.

Hand, B., Norton-Meier, L. A., Gunel, M., & Akkus, R. (2016). Aligning teaching to learning: A 3-year study examining the embedding of language and argumentation into elementary science classrooms. *International Journal of Science and Mathematics Education*, *14*, 847–863.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*, 359–369.

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., . . . Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Educational Evaluation and Policy Analysis*, *27*, 111–132.

Higgins, J. P. T., & Deeks, J. J. (2008). Selecting studies and collecting data. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 151–186). Chichester, UK: John Wiley.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Hill, H. C. (2004). Professional development standards and practices in elementary school mathematics. *The Elementary School Journal*, *104*, 215–231.

Hill, H. C. (2009). Fixing teacher professional development. *Phi Delta Kappan*, *90*, 470–476.

Hill, H. C., Litke, E., & Lynch, K. (2018). Learning lessons from instruction: Descriptive results from an observational study of urban elementary classrooms. *Teachers College Record*, *120*(12), 1–46.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. San Francisco, CA: John Wiley.

Kennedy, M. M. (1999). *Form and substance in in-service teacher education* (Research Monograph No. 13). Arlington, VA: National Science Foundation.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*, 945–980.

Kim, J. S., & Quinn, D. M. (2012, March 2012). *A meta-analysis of K-8 summer reading interventions: The role of socioeconomic status in explaining variation in treatment effects*. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

Kulik, J. A. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say* (P10446.001). Arlington, VA: SRI International.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington, DC: National Center for Special Education Research.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York, NY: Oxford University Press.

Llosa, L., Lee, O., Jiang, F., Haas, A., O'Connor, C., Van Booven, C. D., & Kieffer, M. J. (2016). Impact of a large-scale science intervention focused on

English language learners. *American Educational Research Journal*, *53*, 395–424.

Malouf, D. B., & Taymans, J. M. (2016). Anatomy of an evidence base. *Educational Researcher*, *45*, 454–459.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., & Tal, R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching*, *41*, 1063–1080.

Miles, K. H., Odden, A., Fermanich, M., & Archibald, S. (2004). Inside the black box of school district spending on professional development: Lessons from five urban districts. *Journal of Education Finance*, *30*, 1–26.

Miller, B., Lord, B., & Dorney, J. A. (1994). *Staff development for teachers: A study of configurations and costs in four districts*. Boston, MA: Education Development Center.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), e1000097. doi:10.1371/journal.pmed.1000097

Murphy, R., Penuel, W., Means, B., Korbak, C., Whaley, A., & Allen, J. (2002). *E-DESK: A review of recent evidence on discrete educational software*. Menlo Park, CA: SRI International.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Author.

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

National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. Washington, DC: The National Academies Press.

National Research Council. (2013). *Next Generation Science Standards: For States, by States*. Washington, DC: The National Academies Press. doi:10.17226/18290

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at scale. *Educational Evaluation and Policy Analysis*, *36*, 127–144.

Penuel, W. R., Sun, M., Frank, K. A., & Gallagher, H. A. (2012). Using social network analysis to study how collegial interactions can augment teacher learning from external professional development. *American Journal of Education*, *119*, 103–136.

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*, 207–236.

Presser, A. L., Vahey, P., & Dominguez, X. (2015, March). *Improving mathematics learning by integrating curricular activities with innovative and developmentally appropriate digital apps: Findings from the next generation preschool math evaluation*. Paper presented at the Society for Research on Educational Effectiveness Spring Conference, Washington, DC.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*, 206–230.

Resendez, M., & Azin, M. (2006). *2005 Prentice Hall Science Explorer randomized control trial*. Jackson, WY: PRES Associates.

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.

Russell, S. J., Economopoulos, K., Mokros, J., Kliman, M., Wright, T., Clements, D. H. et al. (2006). *Investigations in Number, Data, and Space. Grade 1*. Glenview, IL: Pearson Scott Foresman.

Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2010). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. *Journal of Research on Educational Effectiveness*, *4*, 1–24.

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

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

Schneider, M. C., & Meyer, J. P. (2012). Investigating the efficacy of a professional development program in formative classroom assessment in middle-school English language arts and mathematics. *Journal of Multidisciplinary Evaluation*, *8*(17), 1–24.

Schwartz-Bloom, R. D., & Halpin, M. J. (2003). Integrating pharmacology topics in high school biology and chemistry classes improves performance. *Journal of Research in Science Teaching*, *40*, 922–938.

Shavelson, R. J., & Towne, L. (Eds.). (2001). *Scientific research in education*. Washington, DC: The National Academies Press.

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

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*, 5–14.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*, 427–515.

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, *79*, 839–911.

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

Smith, M. S., & O'Day, J. (1990). Systemic school reform. *Journal of Education Policy*, *5*, 233-267. doi:10.1080/02680939008549074

Sparks, D. (2002). *Designing powerful professional development for teachers and principals*. Oxford, OH: National Staff Development Council.

Stein, M. K., Remillard, J., & Smith, M. S. (2007). How curriculum influences student learning. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 319–369). New York, NY: Macmillan.

Sussman, J., & Wilson, M. R. (2019). The use and validity of standardized achievement tests for evaluating new curricular interventions in mathematics and science. *American Journal of Evaluation*, *40*(2), 190–213. doi:10.1177/1098214018767313

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations and a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*, 13–30.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*, 85–112.

Taylor, J. A., Getty, S. R., Kowalski, S. M., Wilson, C. D., Carlson, J., & Van Scotter, P. (2015). An efficacy trial of research-based curriculum materials with curriculum-based professional development. *American Educational Research Journal*, *52*, 984–1017.

Taylor, J. A., Kowalski, S. M., Polanin, J. R., Askinas, K., Stuhlsatz, M. A., Wilson, C. D., . . . Wilson, S. J. (2018). Investigating science education effect sizes: Implications for power analyses and programmatic decisions. *AERA Open*, *4*(3). doi:10.1177/2332858418791991

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2008). *Teacher professional learning and development*. Brussels, Belgium: International Academy of Education.

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*, 604–634.

Van Egeren, L. A., Schwarz, C., Gerde, H., Morris, B., Pierce, S., Brophy-Herb, . . . Stoddard, D. (2014, August). *Cluster-randomized trial of the efficacy of early childhood science education with low-income children: Years 1-3*. Poster presented at the 2014 Discovery Research K-12 PI Meeting, Arlington, VA.

What Works Clearinghouse. (2008). *Middle school math curricula*. Washington, DC: U.S. Department of Education.

What Works Clearinghouse. (2010). *Procedures and standards handbook* (Version 2.1). Washington, DC: U.S. Department of Education.

Wilson, S. J., Tanner-Smith, E. E., Lipsey, M. W., Steinka-Fry, K., & Morrison, J. (2011). Dropout prevention and intervention programs: Effects on school completion and dropout among school aged children and youth. *Campbell Systematic Reviews*, *8*. doi:10.4073/csr.2011.8

Wilson, S. M. (2013). Professional development for science teachers. *Science*, *340*, 310–313. doi:10.1126/science.1230725

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

Zaslow, M., Tout, K., Halle, T., Whittaker, J. V., & Lavelle, B. (2010). *Toward the identification of features of effective professional development for early childhood educators, literature review*. Washington, DC: Office of Planning, Evaluation and Policy Development, U.S. Department of Education.

Authors

KATHLEEN LYNCH is a postdoctoral research associate at the Annenberg Institute at Brown University. Her research focuses on education policy

and strategies to reduce educational inequality, particularly in mathematics.

HEATHER C. HILL is the Jerome T. Murphy Professor in Education at the Harvard Graduate School of Education. Her primary work focuses on teacher and teaching quality and the effects of policies and programs aimed at improving both.

KATHRYN E. GONZALEZ is a doctoral candidate at the Harvard Graduate School of Education. Her research focuses on the effects of policies and programs aimed at improving the quality of children's experiences in early childhood education and the impacts of instructional practices and classroom quality on children's early development.

CYNTHIA POLLARD is a doctoral student at the Harvard Graduate School of Education. Her research focuses on characteristics of teachers and teaching that are especially effective for students from historically marginalized groups.

Manuscript received February 26, 2018
First revision received October 16, 2018
Second revision received February 5, 2019
Third revision received March 18, 2019
Accepted March 29, 2019