A Meta-Analysis of the Experimental Evidence Linking STEM Classroom Interventions to Teacher Knowledge, Classroom Instruction, and Student Achievement

Kathryn Gonzalez[1]
Kathleen Lynch[2]
Heather C. Hill[3]

[1]Mathematica
[2]Neag School of Education, University of Connecticut
[3]Harvard Graduate School of Education

Abstract

Despite growing evidence that classroom interventions in science, technology, engineering, and mathematics (STEM) can increase student achievement, there is little evidence regarding how these interventions affect teachers themselves and whether these changes predict student learning. We present results from a meta-analysis of 37 experimental studies of preK-12 STEM professional learning and curricular interventions, seeking to understand how STEM classroom interventions affect teacher knowledge and classroom instruction, and how these impacts relate to intervention impacts on student achievement. Compared with control group teachers, teachers who participated in STEM classroom interventions experienced improvements in content and pedagogical content knowledge and classroom instruction, with a pooled average impact estimate of +0.56 standard deviations. Programs with larger impacts on teacher practice yielded larger effects on student achievement, on average. Findings highlight the positive effects of STEM instructional interventions on teachers, and shed light on potential teacher-level mechanisms via which these programs influence student learning.

Keywords: Professional development, curriculum, mathematics, science, STEM

A Meta-Analysis of the Experimental Evidence Linking STEM Classroom Interventions to Teacher Knowledge, Classroom Instruction, and Student Achievement

Given persistent concerns that the U.S. education system is not adequately preparing students in science, technology, engineering, and mathematics (STEM) (National Research Council, 2013), researchers and policymakers have made significant investments in programs aimed at improving instructional quality in STEM classrooms. Two of the core mechanisms by which policy investments seek to improve the quality of preK-12 STEM teaching in the United States are the provision of teacher professional development and novel curriculum materials. Research syntheses indicate that causal studies of interventions in these categories tend to show mean positive impacts on students' mathematics and science learning, on average (e.g., Blank & De las Alas, 2009; Lynch et al., 2019; Scher & O'Reilly, 2009; Slavin et al., 2009).

Influential logic models describing these instructional improvement interventions posit that beneficial impacts on student learning operate through changes in teachers and teaching. In these models, teachers' participation in professional learning experiences (Cohen & Hill, 1998; Desimone, 2009; Yoon et al. 2007) and implementation and enactment of novel curriculum materials (Ball & Cohen, 1996; Confrey, 2006; Davis et al, 2016) are hypothesized to strengthen teachers' knowledge, skills, and instructional practice. These changes then catalyze improved student learning outcomes (Kennedy, 2016; Scher & O'Reilly, 2009).

In the current research, we conduct a comprehensive meta-analysis of contemporary experimental studies to empirically test the combined elements of this core logic model. The current review is needed for four reasons. First, we include all efforts to improve classroom instruction, whether through teacher professional development or the implementation of novel curriculum materials, expanding upon the literature identified in prior reviews (e.g., Garrett,

2019) and allowing a more comprehensive testing policymakers' logic model. Second, this review examines the effects of preK-12 STEM professional learning and curriculum improvement initiatives on two outcomes, teacher knowledge and classroom instruction, again allowing a more comprehensive test of classroom interventions' logic model than in prior reviews to date. Third, it links teacher outcomes to students' outcomes, enabling the field to connect the dots linking impacts on teachers to changes in student outcomes. Prior reviews typically examine only classroom instruction (e.g., Egert et al., 2018; Kraft et al., 2018); while such analyses are important, they omit an examination of teacher knowledge outcomes despite its theoretical importance. Given that the core purpose of schooling is student learning, the current review thus takes the critical step of investigating how STEM intervention-induced changes in teachers' knowledge and practice may catalyze improvements in student achievement.

Finally, we can identify no prior synthesis investigating how STEM-specific classroom interventions influence teacher knowledge and classroom instruction, and how these changes in turn support student learning. Prior research suggests that the demands of high-quality STEM instruction may differ from the demands on teachers in other subjects (Hill & Lovison, 2021; Pasley et al., 2004). Therefore, syntheses of professional learning interventions focusing on other subject areas or than span across subjects may not provide insight into how classroom interventions operate in STEM contexts. Our analysis enables us to isolate programmatic and contextual features of interventions that predict stronger impacts on teachers' knowledge and instruction specifically in STEM classrooms.

Thus in this review, we synthesize nearly three decades' worth of experimental evidence on the causal impacts of STEM interventions on both teacher knowledge and classroom instruction, and how these outcomes link to student achievement. We employ exclusively

randomized controlled trials, the "gold standard" of evidence supporting causal inference. In doing so, our investigation permits us to evaluate the extent to which the empirical evidence accumulated over this time period supports key components of policymakers' logic model; namely, that improvements in teacher knowledge and instructional practice will lead to improved student outcomes.

Literature Review

Citing concerns about international economic competitiveness, the availability of skilled workers, and, later, academic competitiveness, policymakers in the U.S. have prioritized improving STEM student outcomes for the past several decades (AAAS, 1990; NCEE, 1983; NGA, 2007; NRC, 2007; Welch, 1979). Many of these efforts called for more STEM literacy among U.S. high school and college graduates, with dual attention to meeting employers' needs and fueling innovation in the technology sector. Concerns about workforce readiness dovetailed with new notions of how students learn content (Bransford et al., 2000), with the result that most STEM-related policy efforts indicated a preference for conceptual understanding and application of knowledge over rote memorization and basic understandings of STEM topics.

Achieving these STEM student outcomes, reformers quickly learned, required a coordinated strategy aimed at changing instruction in U.S. classrooms (Welch, 1979). Reformers observed that teachers would need to deploy new instructional techniques focused on inquiry and building students' knowledge of both disciplinary content and methods (National Council of Teachers of Mathematics, 1991). Teachers would also need to present more academically challenging material to students, commensurate with rigorous grade-level standards for student performance (Smith & O'Day, 1990). These anticipated improvements in instruction, in turn, not

only required comprehensive reforms to U.S. governance and assessment systems, but also to the resources teachers use in their daily work (Cohen, Raudenbush, & Ball, 2003).

One key resource was curriculum materials. Since the country's earliest efforts to improve STEM instruction during the Sputnik era (Welsh, 1979), reformers recognized that new forms of teaching could not occur without materials revised to support that teaching. As a result, policymakers invested in curriculum materials intended to enable more rigorous and conceptually focused STEM instruction (Remillard, 2005). Another key resource was teachers' knowledge. Early studies showed that efforts to implement standards-based instruction often derailed in classrooms where teachers lacked sufficient content knowledge to support disciplinary inquiry (Cohen, 1990; Hill et al., 2008). Many teachers had themselves experienced STEM education as the memorization of facts and procedures, the kind of instruction that reformers wished to deemphasize in U.S. classrooms, and struggled when asked to engage students in inquiry, discussion, and disciplinary exploration.

Combined with evidence suggesting that many if not most U.S. teachers lacked such knowledge, particularly in mathematics (An et al., 2004; Ma 1999), and that implementing new instructional practices and curriculum materials would also require significant learning (Ball & Cohen, 1996; Davis & Kracjik, 2005; EEPA, 1990), scholars and practitioners developed STEM-specific professional learning programs. This professional development took a variety of approaches to changing teachers' knowledge, instructional practice, and student outcomes. Some programs focused on building teachers' content knowledge, sometimes largely in isolation from work in classrooms (Garet et al., 2016). At other times, professional development had teachers explore student activities and lessons with an eye toward helping them learn the content they would later be teaching students (Dash et al., 2012; Jacob, Hill & Corey, 2017; Lewis & Perry,

2017). Other programs focused on instructional practices more exclusively, sometimes using case studies, videos, or other depictions of standards-based instructional practices (Heller et al., 2012) and other times by having teachers analyze and try out new curriculum materials (Borman et al., 2008; Clements et al., 2011; Star et al., 2015). Often, programs combined these foci; a prior meta-analysis revealed substantial overlap between programs focusing on curriculum materials and teacher content knowledge (Lynch et al., 2019), and also suggested that many programs integrated other topics, such formative assessment or best practices for teaching emergent bilingual students (Lang et al, 2015; Llosa et al., 2016; Supovitz & Sirinides, 2018).

Evaluations of STEM-specific instructional improvement programs, conducted over a 25-year period, provide an opportunity to test several key elements of reformers' theories about improving teaching and learning. We first ask:

1. What is the causal impact of STEM classroom interventions on teacher knowledge and classroom instruction?

Past evidence suggests that in literatures outside of STEM instructional improvement programs, classroom interventions generate improvements in instructional practice. In a meta-analysis of 40 professional development studies, Garrett et al. (2019) found an average program impact on instructional practice of 0.42 SD, while in a meta-analysis of 36 early childhood in-service professional development studies, Egert et al. (2018) found average program impacts on classroom instruction of roughly 0.68 SD. Examining 60 studies that included teacher coaching as an element of the teacher learning experience, Kraft et al. (2018) found an average effect size of 0.49 SD. However, none of these meta-analyses tested for impacts on teacher knowledge. Further, these meta-analyses largely included studies on literacy and social-emotional learning, leaving questions about STEM-specific programs' impacts on teacher outcomes unanswered.

Understanding these impacts is of particular interest for STEM interventions since, as noted above, some instructional improvement efforts focus on teacher knowledge in hopes of improving instruction (Dash et al., 2012; Jacobs et al., 2007), while others target instructional practice directly (Borman et al., 2008; Penuel et al., 2011).

Next, we examine the links between impacts on teacher outcomes and changes in student achievement:

2. Are program-induced changes in teacher knowledge and classroom instruction linked to improvements in student achievement?

Both Egert et al. (2018) and Kraft et al. (2018) found some evidence of relationships between impacts on classroom instruction and improvements in student achievement in their reviews of early childhood professional development and coaching programs, respectively. Using a subset of included studies that contained information on student outcomes, the authors found positive correlations between gains in instructional quality and gains in student outcomes, though in the Kraft et al. (2018) meta-analysis this relationship was not significant. As neither study tested for impacts on teachers' knowledge, prior authors did not examine the links between impacts on teacher knowledge and student achievement.

Next, we investigate whether intervention characteristics moderate impacts on teacher knowledge and classroom instruction:

3. Do specific features and foci of STEM classroom interventions programs predict positive impacts on teacher knowledge and classroom instruction?

Studies have yielded conflicting evidence regarding the association between various features of instructional improvement interventions and the improvements generated in teacher outcomes. For example, Egert et al. (2018) found that interventions aimed at improving the use of a specific

curriculum did not predict impacts on instruction, while Kraft et al. (2018) found that such a focus did significantly predict outcomes. Similarly, prior meta-analyses examining teacher and student outcomes have yielded mixed findings about links between PD duration and intervention impacts (Egert et al., 2018; Lynch et al., 2019; Yoon et al., 2007). These conflicting findings help motivate our analyses of *intervention features,* examining whether the provision of curriculum materials in combination with professional development, compared to professional development only, moderates program effects on teacher knowledge and practice (which we refer to as intervention type), and our analyses examining whether PD duration is associated with program impacts. We also explore whether *PD focus* moderates the program effects on teacher knowledge and practice, including whether intervention impacts differ based on whether the professional development component of programs focused on teacher knowledge and classroom instruction.

Finally, meta-analysis affords the opportunity to examine what kinds of programs are most effective, for whom, and under what conditions. For example, small-sized programs that offer intensive resources to a limited number of teachers may have larger impacts than scaled-up versions of these programs (Hill, 2004; 2009). Program impacts on instructional quality did appear smaller in the scaled-up programs reported in Kraft et al.'s (2018) coaching meta-analysis, though not in Egert et al.'s (2018) meta-analysis of early childhood in-service programs. Prior work (Lynch et al., 2019) found a marginal trend toward smaller impacts on students from instructional improvement interventions conducted in high-poverty settings. These issues motivate our fourth research question:

4. Do impacts on STEM teacher knowledge or classroom instruction differ by contextual features of the student and teacher sample?

We describe our meta-analytic search procedures and analyses in the next section.

Methods

Defining STEM Classroom Interventions

For the purposes of the current review, we define STEM classroom interventions to include programs that aimed to improve student learning in STEM via teacher professional development, the provision of novel curriculum materials, or both. We excluded interventions that lacked an instructional improvement component, such as afterschool tutoring programs and home-based computerized skills practice, and those that did not involve classroom teachers, such as interventions in which researchers provided all instruction to students directly.

Search and Screening Procedures

We applied the following search procedures to capture relevant published and unpublished experimental studies of STEM classroom interventions produced between 1989 and 2018. We conducted a first comprehensive materials search in 2016; we repeated the search in 2019 to update the pool with additional studies produced through 2018. To be included in the meta-analysis, studies had to meet the following criteria relating to study design, intervention, sample, and outcomes: (1) Include students in grades preK-12, (2) Focus on classroom-level STEM instructional improvement through professional development and/or a change in curriculum materials, (3) Employ a randomized experimental design, (4) Be published in 1989 or later, (5) Be written in English, and (6) Report sufficient data to calculate one or more effect sizes for both teacher and student outcomes. We focus specifically on studies that include both teacher and student outcomes because they are directly aligned with the intended impacts of classroom interventions in our theoretical model.

We searched in several channels. We began by scanning the reference lists of prior research syntheses for studies published between the years 1989 through 2004 that examined the topics of teacher professional development and curriculum improvement in mathematics and science (Blank et al., 2008; Cheung et al., 2017; Furtak et al.,, 2012; Garrett et al., 2018; Gersten et al., 2014; Kennedy, 1999, 2016; Scher & O'Reilly, 2009; Slavin & Lake, 2008; Slavin et al., 2009; Slavin et al., 2014; Timperley et al., 2008; S. M. Wilson, 2013; Yoon et al., 2007; Zaslow et al.,, 2010). Due to resource constraints, along with the low probability of unearthing randomized trials that prior synthesists did not find in their comprehensive searches, we did not conduct further literature searches for materials dated prior to 2004[1]. We next conducted electronic library searches using the databases Academic Search Premier, ERIC, Ed Abstracts, PsycINFO, EconLit, and ProQuest Dissertations & Theses, for the period 2004 to 2018, using subject-related search terms adapted from Yoon et al. (2007) and methodology-related keywords adapted from Kim and Quinn (2013)[2]. We also searched the websites of Regional Education Labs, WWC, the World Bank, Inter-American Development Bank, Empirical Education,

[1] We argue that restricting our search to studies published in 2004 or later is reasonable given that prior reviews have conducted exhaustive searches for research conducted in earlier decades (e.g., Slavin & Lake, 2008; Slavin et al., 2009; Slavin et al, 2014). This approach is also in line with with What Works Clearinghouse protocols, which generally limit their scope to studies published in the past 20 years (Brown et al. 2008).

[2] The search parameters used to conduct electronic database searches are 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"; "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").

Mathematica, MDRC, and American Institutes for Research (AIR), and the abstracts of the Society for Research on Educational Effectiveness (SREE) conference for relevant materials. Lastly, 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 did not search more recent grant abstracts as we assumed that more recent studies may still have been ongoing. We searched electronic databases and the Web to identify relevant studies resulting from the awards, and contacted study PIs to request reports if we could identify no publications with impact results from their grants. These searches yielded 9,214 records from database searches and 1,391 records identified through other sources. After removing duplicates, this yielded 8,785 studies.

Second, raters screened each of the studies' titles and abstracts to identify potentially relevant studies published between 1989 and 2018 that covered grades preK-12, included student outcomes, and focused on math and science-specific content and/or instructional strategies. A total of 780 studies met the initial relevance criteria and were advanced to full-text screening. Third, two authors independently examined the full text of each study and applied more detailed content and methodological criteria listed above, including requiring a randomized experimental research design and reporting impacts on both teacher and student outcomes, then met to reconcile inclusion decisions. See Figure 1 for a PRISMA diagram.

Analytic sample. The final sample includes 37 studies contributing 165 effect sizes for teacher outcomes and 111 effect sizes for student achievement outcomes. These include separate effect sizes for each assessment, treatment contrast, and sample of teachers reported by the study. We also categorized each teacher outcome as either a teacher knowledge (including teacher content knowledge and pedagogical content knowledge) or classroom instruction outcome. For

all analyses, we first considered classroom instruction and teacher knowledge outcomes simultaneously, and then considered each group of outcomes separately. Data limitations in the original studies' outcome reporting preclude us from conducting separate analyses examining impacts on teacher content knowledge as compared with pedagogical content knowledge outcomes. Specifically, too few studies reported impacts on teachers' pedagogical content knowledge ($k = 9$) to permit fitting statistical models with this outcome variable separated from other measures of teacher content knowledge. Prior research suggests that teachers' content knowledge and pedagogical content knowledge capture correlated but distinct constructs (Charalambous et al., 2019; Kleickmann et al., 2015; Krauss et al., 2008), which may influence student learning in different ways (Baumert et al., 2010; Kersting et al., 2012). In the present study, however, we combined these into one category to have a sufficient sample size to examine impacts on teacher knowledge outcomes separately from classroom instruction outcomes.

Study Coding

We developed content codes based on prior meta-analyses and the literature on instructional improvement (e.g., Kennedy, 2016; Kraft et al., 2018; Scher & O'Reilly, 2009). We reviewed prior meta-analyses and systematic reviews of the instructional improvement literature to develop broad categories of codes (e.g., professional development, curriculum) as well as specific codes (e.g., PD focus on content-specific instruction, number of PD contact hours). We coded a sample of studies with an initial set of codes, and then refined the codes as needed. Study authors and trained research assistants coded full-text studies. After establishing interrater reliability at the start of the coding process (i.e., 80% agreement), researchers coded studies in pairs. Each researcher in the pair first coded the study independently, then pairs met to resolve all

coding disagreements through discussion (for more details about code development and study coding, see Lynch et al., 2019).

In the present study, we focused our analyses on three sets of codes closely related to our research questions. First, we focused on two *intervention features* of theoretical interest: whether programs combined PD with curriculum materials or provided PD only, which we refer to as *intervention type,* and PD duration. Growing evidence suggests that PD in combination with specific curriculum can promote student outcomes (Alozie, Moje & Krajcik, 2010; Lynch et al., 2019), although other research has found that classroom interventions centered on curriculum implementation are no more effective than interventions without this emphasis (Egert et al., 2018). At the same time, meta-analyses have yielded conflicting findings about the role of PD duration (Egert et al., 2018; Lynch et al., 2019; Yoon et al., 2007).

Second, we examined the programmatic focus of the professional development component of the intervention, which we refer to as *PD focus*. First, we coded for whether the professional development focused on improving teacher knowledge, including improving teachers' content knowledge or pedagogical content knowledge, and improving teachers' knowledge of how students learn. Second, we coded for whether the professional development focused on improving aspects of classroom instruction, including content-specific instructional strategies, generic instructional strategies, and content-specific formative assessment. These codes allowed us to examine whether programmatic focus on each of these areas, as well as programmatic focus on both teacher knowledge *and* classroom instruction, moderated program impacts on teacher outcomes.

Third, we examined how the impacts of classroom interventions varied based on context and sample, which we refer to as *contextual features.* Specifically, we coded for contextual

features including the teacher sample size, characteristics of the student sample (e.g., the percent of low income or free or reduced-price lunch-eligible students and the percent of emergent bilingual students), student grade level, and the whether the intervention took place in an urban versus suburban or rural school district.

Effect Size Calculation

We calculated standardized mean difference effect sizes for impacts on teacher outcomes and student achievement outcomes using Hedges's *g*:

$$g = J \text{ x} \frac{(\overline{Y_E} - \overline{Y_C})}{S*}$$

Where $\overline{Y_E}$ represents the average treatment group outcome, $\overline{Y_C}$ represents the average control group outcome, $S*$ represents the pooled within-group standard deviation, and J is a correction factor to avoid bias in small samples.

Effect sizes were calculated based on author-reported effect sizes, raw means and standard deviations, and other author-reported results. Effect sizes were calculated using the software package *Comprehensive Meta Analysis* for the majority of cases. Where possible, we calculated effect sizes that were adjusted for covariates (e.g., pretest scores). We used the following decision rules to calculate effect sizes: If authors reported a standardized mean difference effect size (e.g., Cohen's *d*) we converted author-reported effect sizes to Hedges's *g*. If authors did not report a standardized mean difference effect size but reported a covariate-adjusted mean difference (e.g., a coefficient from a regression model) and unadjusted standard deviations, we calculated a standardized mean difference effect size and converted to Hedges's g. If adjusted mean differences were not reported, we calculated effect sizes based on raw posttest means and standard deviations. If this information was not available, effect sizes were

calculated from other results (e.g., results of ANOVAs). We used the same decision rules to calculate effect sizes for teacher outcomes and student outcomes.

Impacts of Classroom Interventions on Teacher Outcomes

We estimated meta-regression models to examine the impacts of classroom interventions on teacher knowledge and classroom instruction outcomes. Many of the studies in our sample yield multiple effect sizes for impacts on teacher knowledge and/or classroom instruction outcomes. Effect sizes nested within the same study are likely to be correlated, such as when one study yields multiple effect sizes for one underlying construct or multiple, related constructs. These dependencies, also referred to as *correlated effects*, violate the assumptions of statistical independence required for the use of traditional meta-analytic methods. Therefore, we use a robust variance estimation (RVE) approach to properly model our data in the presence of correlated effects. The RVE approach, developed by Tanner-Smith & Tipton (2014), adjusts standard errors to account for correlations between multiple effect sizes from the same study. This approach has been widely used in recent meta-analyses (e.g., Clark, Tanner-Smith, & Killingsworth, 2016; Kraft et al., 2018; Lynch et al., 2019). Importantly, this approach allows us to include multiple effect sizes from the same study and avoid the loss of information that would arise from dropping effect sizes or calculating average effect sizes within each study.

In estimating our RVE meta-regression models, we used the inverse variance weights recommended by Tanner-Smith and Tipton (2014). The weight for effect size *i* in study *j* is calculated by the following:

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

Where v_{*j} is the mean of within-study sampling variances (SE_{ij}^2), 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. Effect sizes from studies contributing a larger number of effect sizes and effect sizes from studies with higher sampling variances are given lower weight. We used the *robumeta* package in Stata 16 (developed by Tanner-Smith & Tipton, 2013) to estimate all RVE meta-regression models, and used the recommended default value for the assumed correlation between effect sizes of 0.80 (Tanner-Smith & Tipton, 2014).

We first estimated separate unconditional meta-regression models to estimate the overall impact on all teacher outcomes, including teacher knowledge and classroom instruction outcomes simultaneously. We then estimated separate unconditional meta-regression models to estimate overall impacts on teacher knowledge and classroom instruction outcomes separately. Tests for heterogeneity in effect sizes across studies used in traditional meta-analysis are not available with the RVE approach (Tanner-Smith & Tipton, 2014; Tanner-Smith, Tipton & Polanin, 2016). Instead, we report the method of moments estimate of τ^2, the between-studies variance component, as a measure of between-study heterogeneity in effect sizes.

We then fit additional conditional meta-regression models to examine whether *intervention features* – intervention type and PD duration – moderate program impacts on instructional practice and teacher knowledge. First, we fit conditional models to examine whether impacts differ for interventions that provided PD and new curriculum materials, rather than PD only. Second, we fit additional conditional models to examine whether impacts differ based on PD duration by including the number of PD contact hours as a moderator. As an alternative specification, we also included an indicator for whether the number of PD contact hours was above the sample median (45 hours) as a moderator.

We then fit conditional meta-regression models to examine whether *PD focus,* including program focus on aspects of teacher knowledge, classroom instruction, or both, moderate study impacts on classroom instruction and knowledge outcomes. First, we fit a series of models that included indicators for whether the PD focused on specific aspects of *teacher knowledge*, (teacher content knowledge/pedagogical content knowledge and teacher knowledge of how students learn) as well as indicators for whether the PD focused on specific aspects of *classroom instruction* (generic instructional strategies and content-specific formative assessment). These models controlled for study focus on all four aspects of PD focus simultaneously. Second, we fit a series of models that included an indicator for whether PD focused on at least one aspect of teacher knowledge *and* at least one aspect of classroom instruction, to examine whether interventions that focused on both teacher knowledge and classroom instruction had different impacts on teacher outcomes compared to interventions that focused on only one (or neither) of these areas.

Finally, we fit additional conditional meta-regression models to examine whether impacts differ based on additional characteristics of the studies and contexts in which they were conducted. These characteristics included the size of the teacher sample and various student and school characteristics. Student and school characteristics included student income (percent low-income or eligible for free or reduced-price lunch; whether a majority of students were low-income or eligible for free or reduced-price lunch), student emergent bilingual status, student race/ethnicity, and school district urbanicity (urban vs. suburban or rural). Patterns of missing data varied across these student and school characteristics. Therefore, we estimated separate models that considered each student and school characteristic separately.

All conditional models also featured additional study characteristics as covariates, including whether effect sizes were adjusted for covariates and whether interventions focused on math or science. In some studies, there was within-study variability in moderators or covariates (e.g., contact hours varied across multiple treatment-control contrasts). In these cases, we followed the recommended approach of including the study-level mean value of this moderator (Tanner-Smith & Tipton, 2014).

Linking Impacts on Teacher and Student Outcomes

To link impacts on teacher-level outcomes to impacts on student-level outcomes, we adapted the approaches used in recent meta-analyses (e.g., Egert et al., 2018; Kraft et al., 2018). First, we calculated mean effect sizes for impacts on teacher outcomes and student achievement outcomes for each treatment-control contrast in our sample. In order to examine whether impacts on classroom instruction and teacher knowledge outcomes are separately associated with impacts on student achievement, we calculated three treatment-level mean effect sizes for impacts on teacher outcomes: (1) mean effect sizes for impacts on all teacher outcomes, including both teacher knowledge and classroom instruction outcomes; (2) mean effect sizes for teacher knowledge outcomes; and (3) mean effect sizes for classroom instruction outcomes.

Some studies ($k = 8$) in our sample reported impacts based on multiple treatment-control contrasts. In the presence of within-study, between-treatment variation in teacher and student effect sizes, aggregating effect sizes to the study level could obscure the links between impacts on teacher and student outcomes. Therefore, we calculated mean effect sizes at the treatment level rather than at the study level to more directly test whether programs that supported improvements in instructional practice and teacher knowledge also increased student

achievement. We also confirmed that we obtained similar results with effect sizes that are aggregated at the study level rather than the treatment contrast level.

After calculating treatment-level mean effect sizes for teacher and student outcomes, we then estimated a series of regression models predicting mean impacts on student outcomes as a function of mean impacts on teacher outcomes. We estimated three separate models, including each of the treatment-level mean effect sizes for impacts on teacher outcomes described above as predictors. These models were weighted by the average of the inverse effect size variances for teacher outcomes. These models also featured additional study characteristics as covariates, including whether effect sizes were adjusted for covariates and whether interventions focused on math or science. As in our models examining impacts on teacher outcomes, we included the study-level mean value of covariates in cases where there was within-study variability in the value of covariates.

Results

Study Characteristics

Table 1 provides descriptive information about the study designs and programs in the sample. The sample included 37 studies with 46 treatment-control contrasts. These studies yielded a total of 165 teacher outcomes effect sizes. Of the included studies, 21 (57 percent) featured both professional development and new curriculum materials; 16 studies (43 percent) featured professional development only. A majority of studies focused on math (23 studies; 62 percent) and roughly one third focused on science (12 studies; 32 percent); two studies focused on both math and science (5 percent). Studies included a mix of grade levels, ranging from preschool through high school. Interventions included an average of 56 professional development contact hours; in a majority of studies, intervention activities took place over two or

more semesters (25 studies; 72 percent). As previously noted, all studies in the sample were randomized controlled trials (RCTs).

Table 1 also shows the percentage of studies in which the intervention focused on different aspects of teacher knowledge and classroom instruction. A majority of studies included professional development focused on at least one aspect of teacher knowledge, including improving teacher content knowledge or pedagogical content knowledge (23 studies; 62 percent) and improving teacher knowledge of how students learn (21 studies; 57 percent). The majority of studies also focused on at least one aspect of classroom instruction. Nearly all focused on content-specific instructional strategies (36 studies; 97 percent); therefore, we were unable to examine this feature as a moderator. A smaller proportion of studies focused on content-specific formative assessment (7 studies; 19 percent) and on generic instructional strategies (5 studies; 14 percent). Over two-thirds of studies (26 studies; 70 percent) included a focus on at least one aspect of teacher knowledge *and* at least one aspect of classroom instruction.

The studies in our sample contributed a total of 165 effect sizes representing a mix of impacts on teacher knowledge and classroom instruction outcomes. As shown in Table 1, 51 effect sizes (31 percent) captured impacts on teacher knowledge, including teacher content knowledge or pedagogical content knowledge. These included impacts on math content knowledge, science content knowledge, and pedagogical content knowledge. In addition, 114 of these effect sizes (69 percent) captured impacts on classroom instruction. These included impacts on both observational and self-report measures of instructional practice. Most effect sizes were based on intervenor-developed outcome measures (132 effect sizes; 80 percent), although some

effect sizes were based on standardized (24 effect sizes; 15 percent) or other (9 effect sizes; 6 percent) outcome measures.

Overall Average Impacts on Teacher Outcomes

Table 2 presents the results of estimating unconditional RVE regression models examining study impacts on teacher knowledge and classroom instruction outcomes. Across all included studies, we found an average weighted impact on all teacher outcomes (including teacher knowledge and classroom instruction outcomes) of 0.56 SD ($p < .001$). The prediction interval based on the method-of-moments estimate of the between-study variance provides an indication of between-study heterogeneity in impacts, indicating that true effect sizes would be expected to range from -0.19 SD to 1.32 SD.

When we consider teacher knowledge and classroom instruction outcomes separately, we find studies yielded positive, similarly sized impacts on these two groups of outcomes. Of the 37 studies in our sample, 20 studies contributed at least one effect size for teacher knowledge and 26 studies contributed at least one effect size for classroom instruction. Among the 20 studies with information on teacher knowledge outcomes, we found an average weighted impact on teacher knowledge of 0.53 SD ($p < .001$). Among the 26 studies with information on classroom instruction outcomes, we found an average weighted impact on classroom instruction of 0.57 SD ($p < .001$). Prediction intervals also indicated substantial between-study heterogeneity in impacts leading to questions about the factors that may explain the observed variability.

Linking Impacts on Teacher and Student Outcomes

Table 3 presents the results of estimating weighted regressions that test the associations between intervention impacts on teacher knowledge and classroom instruction outcomes, and student achievement outcomes. When we considered teacher knowledge and classroom

instruction outcomes together, we observed a positive, statistically significant association between treatment-level mean impacts on teacher outcomes and treatment-level mean impacts on student achievement outcomes. As shown in the first column of Table 3, we find that a 1 SD increase in teacher knowledge and instruction outcomes is associated with a 0.21 SD improvement in student achievement.

This association is driven primarily by a positive association between treatment impacts on instructional practice and treatment impacts on student outcomes. As shown in the third column of Table 3, we observe a positive, statistically significant association between mean impacts on classroom instruction and mean impacts on student achievement: a 1 SD increase in classroom instruction yields a 0.27 SD change in student achievement. In contrast, we do not observe a similar, statistically significant association between impacts on teacher knowledge and student achievement. As shown in the second column of Table 3, the association between treatment-level mean impacts on teacher knowledge and treatment-level mean impacts on student achievement is positive, but smaller in magnitude and not statistically significant. Although we cannot make causal inferences about these associations, this suggests that interventions that improved instructional practice also promoted student achievement.

Intervention Characteristics and Contextual Factors that Moderate Program Impacts on Teacher Outcomes

Next, we turn to programmatic and contextual factors that moderate impacts on teacher outcomes. We first examined whether *intervention features,* including teachers' time investments (PD duration) and whether the program combined professional development with new curriculum materials (intervention type), moderated intervention impacts on teacher knowledge and classroom instruction (see Table 4). We did not find a significant difference in impacts on

teacher outcomes between studies that provided both professional development and new curriculum materials, compared to professional development only. We observed these nonsignificant associations when we considered teacher knowledge and classroom instructions outcomes simultaneously and separately. We also observed no statistically significant association between the number of professional development contact hours and teacher outcomes. Moreover, we observed nonsignificant associations between professional development duration and impacts on teacher outcomes when we replaced the continuous measure of PD contact hours with an indicator for whether the number of PD contact hours was above the sample median as a moderator.

Second, we examined whether programmatic focus on different aspects of teacher knowledge and classroom instruction (*PD focus*) was associated with improvements in teacher outcomes. We find some evidence that programmatic focus on improving some aspects of teacher knowledge, as well as a dual focus on improving instructional practice and teacher knowledge, moderated intervention impacts on classroom instruction (see Table 5). Interventions that focused on improving teacher knowledge of how students learn STEM content had larger impacts on classroom instruction compared to interventions without this focus (a difference of 0.62 SD; $p < .05$). We also find that interventions that focused on improving at least one area of instructional practice (including content-specific instruction, generic instructional strategies, and/or content-specific formative assessment) *and* on improving teacher knowledge (including content knowledge or pedagogical content knowledge and/or knowledge of how students learn) had somewhat larger impacts on teacher practice than interventions that included a focus on classroom instruction or teacher knowledge only (a difference of 0.33 SD; $p < .10$). However, programmatic focus on specific aspects of classroom instruction was not associated with

statistically significant differences in intervention impacts. Moreover, we do not find evidence that programmatic focus significantly moderated impacts on teacher knowledge outcomes.

We next examined whether other contextual study features moderated intervention impacts on teacher outcomes. We find some evidence that studies with smaller teacher sample sizes had larger impacts on teacher outcomes, although this finding is only marginally significant (see Table S1; online only). We did not observe significant relationships between other contextual features that we coded and intervention impacts. Impacts on teacher outcomes did not differ significantly based on a variety of characteristics of the student sample, including the percentage of students who were low income or eligible for free or reduced-price lunch, the percentage of emergent bilingual students, and the percentage of white students. Impacts on teacher outcomes also did not vary significantly between studies implemented in urban school districts compared to rural or suburban districts (see Table S2; online only), and in different grade levels (see Table S3; online only).

Publication Bias

In all systematic reviews there exists the possibility of publication bias among available studies. We take three approaches to examine this issue in the present sample of studies. We first examine funnel plots to explore whether there is visual evidence of publication bias. We examine three funnel plots, including plots for classroom instruction and teacher knowledge outcomes simultaneously, teacher knowledge outcomes only, and classroom instruction outcomes only. We observe asymmetry in all three funnel plots, providing visual evidence of possible publication bias (see Figures S1 to S3; online only). This pattern appears somewhat more pronounced for teacher knowledge outcomes relative to classroom instruction outcomes.

We then conduct two statistical tests for publication bias. First, we test for publication bias using Egger's regression test. For this test, we first aggregate effect sizes and effect size standard errors to the study level by calculating the average effect size and average effect size standard error, across all effect sizes in each study. We then regress the standard normal deviation (the effect size divided by its standard error) on the inverse of the effect size standard error. This approach tests the null hypothesis that the regression intercept is zero (i.e., that there is not publication bias). If the null hypothesis is rejected this indicates evidence of publication bias (e.g., Egger et al., 1997; Sterne & Egger, 2005). Second, we used a modification of the Egger's test by adding the standard error of the effect sizes as a moderator to the unconditional RVE meta-regression model. If the standard errors of the effect sizes predict the magnitude of the effect sizes, this similarly indicates evidence of publication bias. For both, we conducted this test separately for teacher knowledge and classroom instruction outcomes, teacher knowledge outcomes only, and classroom instruction outcomes only. Results of both of these tests are consistent with the presence of publication bias, and suggest that this may be more pronounced among studies that examined impacts on teacher knowledge (see Tables S4 and S5; online only). These findings are consistent with a recent meta-analysis examining the impact of professional development on instructional practice (Garrett et al., 2019), which also found evidence of publication bias in studies examining the impact of professional development on instruction, and point toward the importance of searching the grey literature when capturing research in this domain.

Sensitivity Checks

Overall average impacts on teacher outcomes. We conducted several sensitivity checks to test the robustness of our findings to different sample and model specifications. Effect

sizes for impacts on classroom instruction represented a combination of self-report and observational measures of instructional practice. To determine whether this mix of self-report and observational outcomes could influence our estimate of the overall average impact on classroom instruction, we first replicated our unconditional RVE meta-regression model after restricting the sample first to effect sizes based on self-report measures of instructional practice, and then to effect sizes based on observational measures of classroom instruction. Results indicate that the overall average impacts were similar for self-report and observational measures of instructional practice (see Table S6; online only).

Effect sizes also represent a mix of standardized, intervenor-developed, and other types of outcome measures. Therefore, we also replicated our unconditional model after restricting the sample to effect sizes based on standardized or intervenor-developed outcome measures, and after restricting the sample to effect sizes based on intervenor-developed outcome measures. We could not examine impacts on standardized or other outcome measures separately due to the small number of effect sizes in each category. Results indicate that excluding these different groups of outcome measures did not substantially affect the magnitude of estimated impacts on classroom instruction or teacher knowledge (see Table S7; online only).

Linking impacts on teacher and student outcomes. We also examined the sensitivity of our findings regarding the links between impacts on teacher and student outcomes to model specification. We confirmed that we received similar results when we examined unweighted associations between intervention impacts on teacher and student outcomes (see Table S8; online only), and when we examined weighted associations between impacts on teacher and student outcomes that used mean effect sizes aggregated to the study level rather than to the treatment-contrast level (see Table S9; online only).

Discussion

In sum, we comprehensively reviewed the rigorous empirical research spanning nearly three decades on the causal impacts of STEM interventions on both teacher knowledge and classroom instruction, and linked these outcomes to student achievement. To what extent did the research evidence support core elements of the major logic model of STEM instructional reform -- namely, that interventions will strengthen teacher knowledge and instructional practice, which will lead to improved student outcomes?

Overall Impacts on Teachers' Knowledge and Instruction, and Links to Student Achievement

The cumulative experimental evidence indicates that classroom STEM interventions had positive impacts on *teacher knowledge*, a key component of reformers' theory of action (e.g., NRC, 2007; 2012). We can identify no prior study that has synthesized the rigorous experimental evidence on this issue. The average weighted impact on teacher knowledge was 0.53 SD. To contextualize the magnitude of this effect, a typical treatment group teacher would be expected to rank approximately 20 percentile points higher than a typical control group teacher on mean indicators of teacher knowledge (Lipsey et al., 2012).

The evidence also shows positive impacts of classroom STEM interventions on *classroom instruction*. We found an average weighted impact on classroom instruction of 0.57 SD. Expressed in terms of percentile ranks, a typical teacher in the treatment group would be expected to rank approximately 22 percentile points higher than a typical control group teacher on measures of instruction (Lipsey et al., 2012). This result comports with prior studies' findings (e.g., Garrett et al., 2019; Kraft et al., 2018). Based on reviewing classroom practice-directed interventions with observation outcomes across content areas, Garrett et al. (2019) found a

pooled mean effect size of 0.42 SD on observation scores. Examining the impacts of teacher coaching interventions, mostly in literacy and general instructional pedagogy, Kraft et al. (2018) found a mean pooled effect size estimate of 0.49 SD on observation score outcomes. The magnitude of these estimates implies that classroom interventions of the type evaluated make a marked difference to both teachers' knowledge and classroom instruction in STEM.

Our data do not permit causal inference about the relative importance of knowledge versus practice emphases in classroom STEM interventions on student outcomes; rather, the evidence is correlational in nature and thus suggestive of links worth noting. With this in mind, the evidence from our analyses connecting teacher and student impacts provides partial support for reformers' theory of action regarding the links between teacher knowledge improvements, instructional practice improvements, and student outcomes. We find supportive evidence consistent with the notion that on average, classroom interventions that have stronger causal impacts on instruction have stronger impacts on students' mathematics and science achievement. On average, a 1 SD improvement in classroom instruction predicted a positive 0.27 SD difference in student test scores – the equivalent of an improvement in student achievement from the 50^{th} to the 61^{st} percentile. Using recently proposed effect size benchmarks for education research (Kraft, 2020), this constitutes a "large" effect.

On the other hand, our models could not confirm the link between improved teacher knowledge and improved student outcomes. Although the coefficient for this association was positive, it was not statistically significant. It is possible that programs that focused on specific aspects of teachers' knowledge may have been more effective than others, such as pedagogical content knowledge versus content knowledge, but the pool of study reports did not describe the types of teacher knowledge emphasized in enough clarity to permit investigation of this issue.

In addition, too few studies examined impacts on both content knowledge and pedagogical content knowledge to compare whether impacts on content knowledge versus PCK had stronger impacts on student outcomes. The overall findings in this domain may imply one of two options. On the one hand, this finding could be due to the relatively small number of studies that provided the necessary information. On the other hand, the pattern of findings is consistent with a scenario in which teacher knowledge improvements may be less influential for strengthening student achievement as compared with improvements in teacher practice.

If so, this finding suggests that several decades of emphasis on improving teachers' content knowledge may have done little to improve student outcomes. Landmark reports, such as *Rising Above the Gathering Storm* (NASEM, 2007), argued that all K–12 science and mathematics teachers should be provided with "high-quality continuing professional development opportunities—specifically those that emphasize rigorous content education" (p. 120). Through both No Child Left Behind's Title II Math-Science Partnerships (MSPs) and funding initiatives sponsored by the National Science Foundation, the federal government spent roughly $2 billion (Hill et al., 2011) funding scientists, mathematicians, and teacher educators to develop teacher learning opportunities that were focused on STEM subject matter, presented via modalities including problem-solving, investigations, and lectures. One possibility raised by our results – and demonstrated by programs primarily aimed at improving content knowledge (e.g., Garet et al., 2010) – is that improving knowledge is not in itself sufficient to improve instructional practice and student outcomes.

As a practical matter, we note that many contemporary STEM professional learning programs eschew an "either/or" approach to improving knowledge and instructional practice, and instead emphasize both of these levers, providing teachers opportunities to deepen their learning

of the content to be taught while also attending to pedagogical practices that teachers will use to portray the content. For instance, when teachers learn to use new curriculum materials they may also be learning the specific subject-matter knowledge embedded in those materials (Remillard & Kim, 2017). Yet, the optimal balance of focus on subject matter knowledge and practice in teacher professional learning remains an important unanswered question in STEM education research. In either scenario, the observed pattern of findings supports the important role of influencing practice alongside knowledge in teacher professional learning and curriculum interventions.

Did Program Features Predict Larger Effects on Teachers' Knowledge and Practice?

The mean positive impacts we observe on teacher outcomes, combined with evidence on between-study heterogeneity in these effects, lead to questions about why some interventions were more effective than others. With respect to instructional practice, interventions that included a focus on how students learn had larger impacts, on average, as compared with interventions that lacked this feature. This finding aligns with an influential body of research in mathematics education that focused on cognitively guided instruction (CGI), drawing teachers' attention to their students' ways of thinking as a means to improve instructional quality (e.g., Carpenter et al., 1989). Interventions that included a combined focus on both teacher knowledge and practice also tended to have stronger impacts on teacher practice outcomes, on average, as compared to interventions that focused on teacher knowledge or practice alone.

Meanwhile, variability in PD impacts on teacher knowledge was not significantly explained by our key moderators, including intervention type, PD duration, or PD focus. We also note the lack of significant relationships between some of our hypothesized moderators and teacher knowledge and practice impacts. We found no significant relationship between

professional development duration and teacher outcomes, including after parsing the data for potential nonlinear relationships. This finding echoes the findings of Kennedy (2016) and Lynch et al. (2019), which did not find a clear benefit of longer-duration professional learning experiences. Speculatively, one possibility is that perhaps shorter interventions focused on more targeted skills, resulting in larger effect sizes on those outcomes. A related possibility is that professional development content and quality may have mattered more than sheer contact hours for supporting teachers' knowledge and instructional practice improvement.

Conclusions and Future Research Directions

The limitations of this study point to promising avenues for future research. First, empirical studies often reported impacts on teacher knowledge or instructional practice, but not both. We urge future research studies to report both types of outcomes, consistent with major logic models of teacher professional learning and to enable future synthesists to empirically test these logic models with larger data pools. In addition, future studies that experimentally vary the amount of emphasis placed in a teacher professional learning experience on strengthening teacher content knowledge versus practice could shed further light on the relative influence of these professional learning emphases on student outcomes.

Second, our dataset afforded us a unique opportunity to connect the dots between intervention-induced improvements in teachers' knowledge and practices and student learning outcomes. The observed associations between teacher and student outcomes cannot, however, be interpreted within a causal framework. As Kraft et al. (2018) have noted, with access to original study data, researchers could shed light on this issue by using random assignment as an instrument to analyze the causal effects of changes in teachers' practices or knowledge on

student learning (e.g., Star et al., 2015) (Murnane & Willett, 2011). Such investigations would be a useful component in future experimental work.

Missing data in original study reports is a perennial challenge for research reviews, and the current synthesis was no exception. One form of missing data relates to the kinds of interventions that are studied via randomized trials and made publicly available by study authors. Researchers have noted that perceptions of the effectiveness of particular genres of interventions may be skewed if the research literature tends to focus disproportionately on "boutique" programs, such as researcher-designed programs that require intensive resource investments to implement, which likely differ from the kinds of programs typically available to teachers in districts (Hill, 2004). Nevertheless, we concur with calls for the importance of conducting more rigorous evaluations of professional learning and curriculum programs that are in widespread use.

Another form of missing data relates to publication bias, or the possibility that studies with null results are more likely to be unavailable through conventional searches. Our extensive search of the grey literature, combined with better reporting practices in the field (e.g., interim and final reports posted on study websites) was intended to mitigate this concern. However, tests for publication bias revealed that our data is consistent with such bias, and thus we must urge caution in the interpretation of our results.

Information about the contexts in which studies are conducted was frequently unreported. Although we had originally hoped to code studies for several features of the school and district context, such as administrative and political support for the intervention as well as the resource indicators including cost per pupil (e.g., Penuel et al., 2010; Wilson, 2013), in most studies the

school and district context was largely a "black box," precluding analysis of these variables. We urge researchers to include information on these contextual variables in future study reports.

Despite the noted limitations, we were able to synthesize nearly three decades' worth of rigorous causal research on the impacts of STEM instructional improvement programs on teachers' knowledge and classroom instruction, and draw connections between teacher impacts and student learning outcomes. Across rigorously evaluated programs, we find that STEM professional learning and curricular interventions had significant and sizable impacts on measures of both teachers' knowledge and their classroom instruction. In turn, improvements in instruction were significant predictors of improvements in student learning. The combined findings point toward the need for future experimental research that builds on the current review and extends the existing evidence base, with the goal of expanding our understanding of innovations that strengthen STEM learning and broaden opportunities for all learners.

References

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*(3), 395-427. https://doi.org/10.1002/sce.20365

American Association for the Advancement of Science. (1990). *Science for all Americans*. New York, NY: Oxford University Press.

An, S., Kulm, G., & Wu, Z. (2004). The pedagogical content knowledge of middle school, mathematics teachers in China and the US. *Journal of Mathematics Teacher Education*, *7*(2), 145-172.

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. https://doi.org/10.3102/0013189X025009006

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133-180. https://doi.org/10.3102/0002831209345157

Blank, R. K., & De las Alas, N. (2009). *The Effects of Teacher Professional Development on Gains in Student Achievement: How Meta Analysis Provides Scientific Evidence Useful to Education Leaders*. Council of Chief State School Officers. One Massachusetts Avenue NW Suite 700, Washington, DC 20001.

Bransford, John D., Ann L. Brown, and Rodney R. Cocking. *How people learn.* Vol. 11. Washington, DC: National Academy Press, 2000.

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.

Charalambous, C. Y., Hill, H. C., Chin, M. J., & McGinn, D. (2019). Mathematical content knowledge and knowledge for teaching: exploring their distinguishability and contribution to student learning. *Journal of Mathematics Teacher Education*, 1-35.

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. https://doi.org/10.1002/tea.21338

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*(1), 79-122. https://doi.org/10.3102/0034654315582065

Cohen, D. K., & Hill, H. C. (1998). Instructional policy and classroom performance: The mathematics reform in California.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119-142. https://doi.org/10.3102/01623737025002119

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, *34*(3), 3-14. https://doi.org/10.3102/0013189X034003003

Desimone, L. M. (2009). Improving impact studies of teachers' professional development:

Toward better conceptualizations and measures. *Educational Researcher*, 38, 181–199. https://doi.org/10.3102/0013189X08331140

EEPA. 1990. Educational Evaluation and Policy Analysis [Whole issue] 12(3):233–353

Egert, F., Fukkink, R. G., & Eckhardt, A. G. (2018). Impact of in-service professional development programs for early childhood teachers on quality ratings and child outcomes: A meta-analysis. *Review of Educational Research*, *88*(3), 401-433. https://doi.org/10.3102/0034654317751918

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. https://doi.org/10.1136/bmj.315.7109.629

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A metaanalysis. *Review of Educational Research*, 82, 300–329. https://doi.org/10.3102/0034654312457206

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915-945. https://doi.org/10.3102/00028312038004915

Garrett, R., Citkowicz, M., & Williams, R. (2019). How responsive is a teacher's classroom practice to intervention? A meta-analysis of randomized field studies. *Review of Research in Education*, *43*(1), 106-137. https://doi.org/10.3102/0091732X19830634

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

Hill, H. C. (2004). Professional development standards and practices in elementary school mathematics. *The Elementary School Journal*, *104*(3), 215-231.

Hill, H. C. (2009). Fixing teacher professional development. *Phi Delta Kappan*, *90*(7), 470-476.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831.

Hill, H. C., Blazar, D., & Lynch, K. (2015). Resources for teaching: Examining personal and institutional predictors of high-quality instruction. *AERA Open*, *1*(4). https://doi.org/10.1177/2332858415617703

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430-511. https://doi.org/10.1080/07370000802177235

Hill, H.C. and Lovison, V. (2021). U.S. Middle School Mathematics Instruction, 2016.

Hill, H. C., Lovison, V., & Kelley-Kemple, T. (2019). Mathematics Teacher and Curriculum Quality, 2005 and 2016. *AERA Open*, *5*(4). https://doi.org/10.1177/2332858419880521

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*(4), 945-980. https://doi.org/10.3102/0034654315626800

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos

predict teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568-589. https://doi.org/10.3102/0002831212437853

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.

Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., ... & Baumert, J. (2015). Content knowledge and pedagogical content knowledge in Taiwanese and German mathematics teachers. *Teaching and Teacher Education*, *46*, 115-126. https://doi.org/10.1016/j.tate.2014.11.004

Kloser, M. (2014). Identifying a core set of science teaching practices: A delphi expert panel approach. *Journal of Research in Science Teaching*, *51*(9), 1185-1217. https://doi.org/10.1002/tea.21171

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547-588. https://doi.org/10.3102/0034654318759268

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*(3), 716. https://doi.org/10.1037/0022-0663.100.3.716

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260-293.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Routledge.

National Commission on Excellence in Education. (1983). *A Nation at Risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

National Governors Association. (2007). *Innovation America: A final report*. Retrieved from www.nga. org/files/live/sites/NGA/files/pdf/0707INNOvATIONFINAL.PDF

National Research Council. (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing?*. National Academies Press.

National Research Council Committee on Prospering in the Global Economy of the 21st Century. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: National Academies Press.

Pasley, J. D., Weiss, I. R., Shimkus, E. S., & Smith, P. S. (2004). Looking Inside the Classroom: Science Teaching in the United States. *Science Educator*, *13*(1), 1-12.

Penuel, W. R., Riel, M., Joshi, A., Pearlman, L., Kim, C. M., & Frank, K. A. (2010). The alignment of the informal and formal organizational supports for reform: Implications for improving teaching in schools. *Educational Administration Quarterly*, *46*(1), 57-95. https://doi.org/10.1177/1094670509353180

Porter, A. C., Garet, M. S., Desimone, L. M., & Birman, B. F. (2003). Providing effective professional development: Lessons from the Eisenhower program. *Science Educator*, *12*(1), 23.

Porter, A. C., Garet, M. S., Desimone, L., Yoon, K. S., & Birman, B. F. (2000). Does professional development change teaching practice? Results from a three-year study.

Remillard, J., & Kim, O. K. (2017). Knowledge of curriculum embedded mathematics: Exploring a critical domain of teaching. *Educational Studies in Mathematics*, *96*(1), 65-81.

Scher, L., & O'Reilly, F. (2009). Professional development for K–12 math and science teachers: What do we really know?. *Journal of Research on Educational Effectiveness*, *2*(3), 209-249. https://doi.org/10.1080/19345740802641527

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*(3), 427-515. https://doi.org/10.3102/0034654308317473

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*(2), 839-911. https://doi.org/10.3102/0034654308330968

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

Smithson, J., & Blank, R. (2006). Indicators of quality of teacher professional development and instructional change using data from surveys of enacted curriculum: Findings from NSF MSP-RETA project.

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. https://doi.org/10.1002/tea.21139

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13-30. https://doi.org/10.1002/jrsm.1091

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

Welch, W. W. (1979). Chapter 7: Twenty years of science curriculum development: A look back. *Review of Research in Education*, *7*(1), 282-306. https://doi.org/10.3102/0091732X007001282

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

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033. *Regional Educational Laboratory Southwest (NJ1)*.

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.

Tables and Figures

Table 1

Sample Sizes and Study Characteristics

Sample sizes		
Total number of studies (treatment-control contrasts)	37 (46)	
Total number of teacher outcomes effect sizes	165	
Study characteristics	*Number of studies*	*Percent of studies or mean (SD)*
Intervention type[a]		
Professional development only	16	43.2%
Professional development + New curriculum materials	21	56.8%
Subject matter focus		
Mathematics only	23	62.2%
Science only	12	32.4%
Mathematics + Science	2	5.4%
Grade level[b]		
Preschool	7	18.9%
Kindergarten	2	5.4%
Early Elementary	8	21.6%
Upper Elementary	15	40.5%
Middle School	12	32.4%
High School	2	5.4%
Professional development hours[c]		56.4 (49.0)
Professional development focus: Teacher knowledge		
Improve content knowledge/pedagogical content knowledge	23	62.2%
Improve knowledge of how students learn	21	56.8%
Professional development focus: Classroom instruction		
Content-specific instructional strategies	36	97.3%
Generic instructional strategies	5	13.5%
Content-specific formative assessment	7	18.9%
Professional development focus: Classroom instruction and teacher knowledge together	26	70.3%
Effect size characteristics	*Number of effect sizes*	*Percent of effect sizes*
Outcome type		
Teacher content knowledge or pedagogical content knowledge	51	30.9%
Classroom instruction	114	69.1%
Effect size adjusted for covariates	107	64.9%
Outcome measure type		
Standardized	24	14.6%
Intervenor-developed	132	80.0%
Other	9	5.5%

[a] Studies with at least one treatment arm that provided new curriculum materials and professional development were included in "Professional development + New curriculum materials."
[b] Studies may have included multiple grade levels.
[c] If professional development hours varied across treatment arms, we calculated the study average.
[d] Includes only studies that provided new curriculum materials. If the percent of lessons replaced with new curriculum materials varied across treatment arms, we calculated the study average.

Table 2

Results of Estimating Unconditional RVE Meta-regression Models Examining the Impacts of Classroom Interventions on Teacher Knowledge and Classroom Instruction

	(1) All teacher outcomes	(2) Teacher knowledge outcomes	(3) Classroom instruction outcomes
Intercept	0.561***	0.531***	0.571***
	(0.062)	(0.079)	(0.083)
N effect sizes	165	51	114
N studies	37	20	26
τ^{2}[a]	0.148	0.108	0.173
.95 prediction interval[b]	[-0.193, 1.315]	[-0.113, 1.175]	[-0.244, 1.386]

Notes: Standard errors in parentheses. All models were estimated using the *robumeta* package in Stata 16 (Tanner-Smith & Tipton, 2015).

+ $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$.

[a] τ^2 is the method-of-moments estimate of the between-study variance component, and was obtained using *robumeta.*

[b] The prediction interval was calculated as: $\hat{\mu} \pm 1.96 * \hat{\tau}$, where $\hat{\mu}$ is the estimated average effect size and $\hat{\tau}$ is the square root of the method-of-moments estimate of the between-study variance component.

Table 3

Results of Estimating Weighted Regression Models Examining the Associations Between Treatment Impacts on Teacher Outcomes and Student Outcomes

	(1)	(2)	(3)
		Student achievement	
All teacher outcomes	0.211*		
	(0.090)		
Teacher knowledge		0.080	
		(0.139)	
Classroom instruction			0.270**
			(0.092)
Controls for study covariates	Yes	Yes	Yes
N treatment arms	46	32	25
N studies	37	26	20

Notes: Standard errors in parentheses. Table presents results of estimating regression models predicting intervention (treatment arm) mean impacts on student outcomes as a function of intervention (treatment arm) mean impacts on teacher outcomes, weighted by the average inverse effect size effect for teacher outcomes. Study covariates include whether intervention focused on math or math/science (vs. science only) and whether effect sizes were adjusted for covariates. Study-level mean values of all covariates were included in the model.

+ $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$.

Table 4

Results of Estimating Conditional RVE Meta-regression Models Examining the Impacts of Classroom Interventions on Teacher Knowledge and Classroom Instruction, including Intervention Type and Dosage as Moderators

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	All teacher outcomes			Teacher knowledge outcomes			Classroom instruction outcomes		
Intervention type:									
Both PD + curriculum	-0.096			-0.326			0.057		
	(0.124)			(0.236)			(0.153)		
PD contact hours[a]		0.018			0.008			0.019	
		(0.013)			(0.021)			(0.016)	
PD contact hours: Above sample median (>45 hours)			0.199			0.123			0.277
			(0.126)			(0.124)			(0.210)
N effect sizes	165	163	163	51	51	51	114	112	112
N studies	37	36	36	20	20	20	26	25	25
Controls for study covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors in parentheses. Study covariates include whether intervention focused on math or math/science (vs. science only) and whether effect sizes were adjusted for covariates. Study-level mean values of all covariates were included in the model. Information on PD contact hours was missing for one study. All models were estimated using the *robumeta* package in Stata 16 (Tanner-Smith & Tipton, 2015).

+ $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$.

[a] PD contact hours measured as PD contact hours/10. PD contact hours was missing for one study in our sample.

Table 5

Results of Estimating Conditional RVE Meta-regression Models Examining the Impacts of Classroom Interventions on Teacher Knowledge and Classroom Instruction, including Professional Development Focus as Moderators

	(1)	(2)	(3)	(4)	(5)	(6)
	All teacher outcomes		Teacher knowledge outcomes		Classroom instruction outcomes	
PD focus: Teacher knowledge						
Improve content knowledge/PCK	0.049		0.358		-0.099	
	(0.188)		(0.387)		(0.143)	
Improve knowledge of how students learn	0.279		-0.207		0.616*	
	(0.195)		(0.334)		(0.205)	
PD focus: Classroom instruction						
Generic instructional strategies	-0.068		-0.104		-0.085	
	(0.173)		(0.236)		(0.270)	
Content-specific formative assessment	0.165		0.074		0.118	
	(0.113)		(0.169)		(0.131)	
PD focus: Classroom instruction and teacher knowledge		0.144		-0.175		0.326+
		(0.126)		(0.168)		(0.169)
N effect sizes	165	165	51	51	114	114
N studies	37	37	20	20	26	26
Controls for study covariates	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors in parentheses. Study covariates include whether intervention focused on math or math/science (vs. science only) and whether effect sizes were adjusted for covariates. Study-level mean values of all covariates were included in the model. Information on PD contact hours was missing for one study. All models were estimated using the *robumeta* package in Stata 16 (Tanner-Smith & Tipton, 2015).

+ $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$.

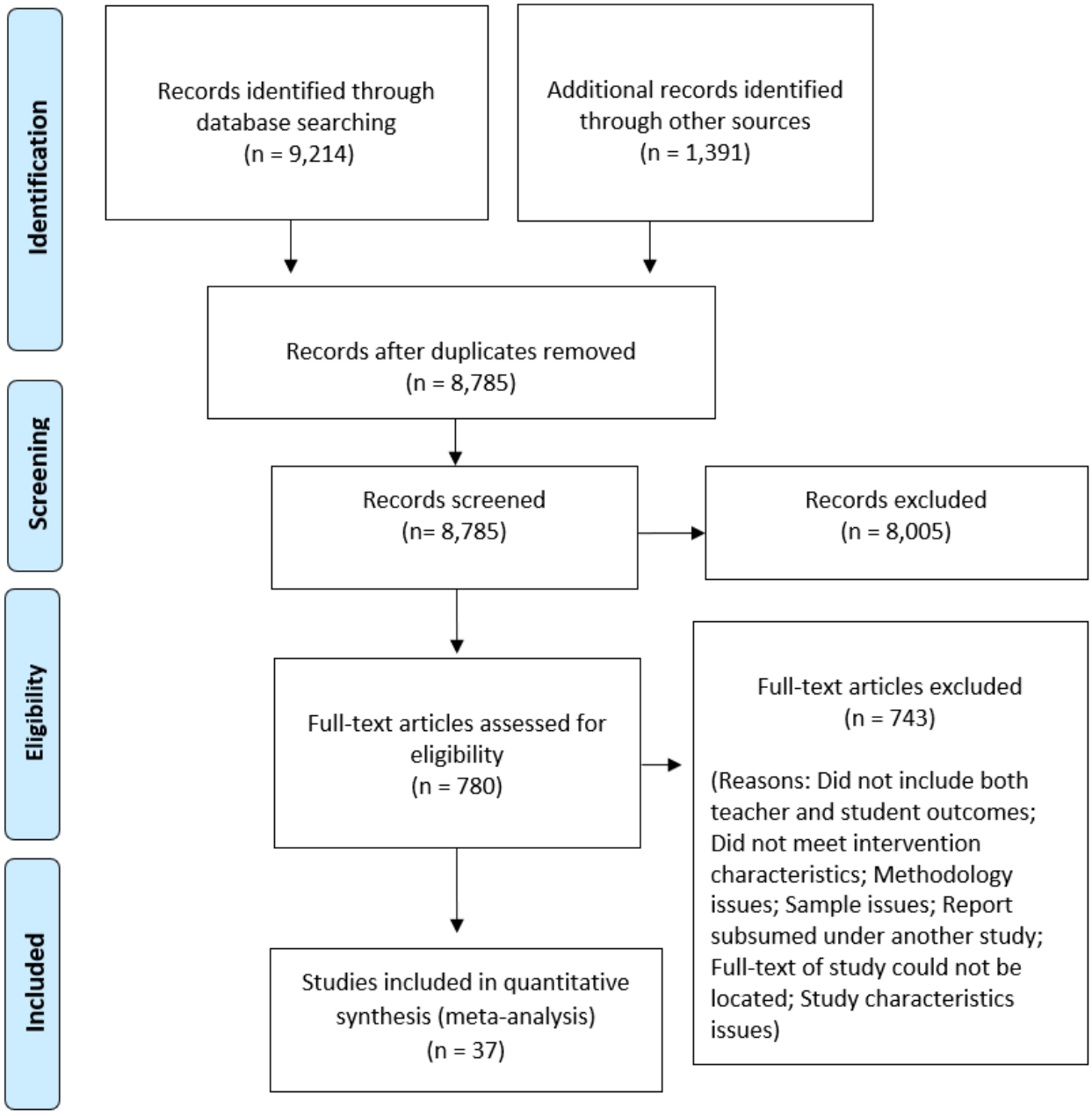

Figure 1
PRISMA Study Screening Flowchart

Source: Moher, Liberati, Tetzlaff, Altman, and The PRIMSA Group (2009)
Note: PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses. Portions of this flowchart adapted from Figure 1 of Authors (2019).