Chapter 4

How Responsive Is a Teacher's Classroom Practice to Intervention? A Meta-Analysis of Randomized Field Studies

Rachel Garrett
Martyna Citkowicz
Ryan Williams
American Institutes for Research

While teacher effectiveness has been a particular focus of federal education policy, and districts allocate significant resources toward professional development for teachers, these efforts are guided by an unexplored assumption that classroom practice can be improved through intervention. Yet even assuming classroom practice is responsive, little information is available to inform stakeholder expectations about how much classroom practice may change through intervention, or whether particular aspects of classroom practice are more amenable to improvement. Moreover, a growing body of rigorous research evaluating programs with a focus on improving classroom practice provides a new opportunity to explore factors associated with changes in classroom practice, such as intervention, study sample, or contextual features. This study examines the question of responsiveness by conducting a meta-analysis of randomized experiments of interventions directed at classroom practice. Our empirical findings indicate that multiple dimensions of classroom practice improve meaningfully through classroom practice-directed intervention, on average, but also find substantial heterogeneity in the effects. Implications for practice and research are discussed.

Policymakers, practitioners, and researchers have all recognized the salience of having highly effective teachers in all of our nation's classrooms. Research repeatedly has demonstrated the importance of teacher quality for student achievement, beyond other school-level characteristics (Aaronson, Barrow, & Sander, 2007; Goldhaber, 2002; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). The influence of teachers

Review of Research in Education
March 2019, Vol. 43, pp. 106–137
DOI: 10.3102/0091732X19830634

can persist over time, with research linking teaching quality in elementary and middle schools to college attendance (Chamberlin, 2013).

Federal education policy has had a long-standing interest in teacher effectiveness, but it has put a particular focus on effective teachers over the past 15 years. This began with the requirements for highly qualified teachers in every school, as mandated by the No Child Left Behind Act in 2002. This was followed by the federal Race to the Top competition and the No Child Left Behind waivers, which both continued to place significant emphasis on ensuring effective instruction, primarily through a focus on more robust teacher evaluation systems. More recently, federal accountability requirements for teacher evaluation have been loosened in the Every Student Succeeds Act reauthorization, but states and districts continue to place a strong emphasis on teaching quality and to allocate significant resources toward professional development for teachers (Jacob & McGovern, 2015).

Alongside the increased focus on teacher effectiveness in education policy, the field of education research has witnessed a substantial increase in rigorous education research in instruction. As shown in Figure 1, a search of the literature for randomized controlled field studies that include a focus on developing classroom practice demonstrates this uptick. The research likely has responded to both the policy impetus and the related need from the field to understand how to support and promote effective instruction through professional learning programs for teachers.

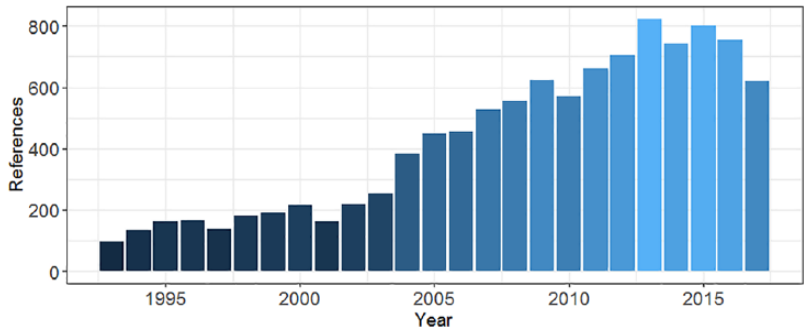
Yet despite the empirical evidence that teachers play an important role in student learning, and the volume of research on interventions that aim to improve classroom practice, an underlying, fundamental question has not yet been addressed: To what extent is a teacher's classroom practice responsive to intervention?

The purpose of this study is to examine the question of responsiveness by conducting a meta-analysis of randomized experiments of interventions directed at classroom practice. We consider how changes in classroom practice may vary by specific aspects of practice, how heterogeneity of effects may relate to features of the interventions, and we seek to identify if there are particularly effective approaches to teacher professional learning. Through this work, we address the following set of questions, which have implications for future design of teacher professional learning programs, and for study design:

1. How does a teacher's classroom practice respond to intervention?
2. Are specific aspects of classroom practice more or less responsive?
3. Are particular intervention features (e.g., coaching, video and technology components, intervention length) associated with improvements in classroom practice?

We find that classroom practice is responsive, and interventions directed toward classroom practice are, on average, able to have meaningful and positive impacts. However, we also find substantial heterogeneity in effects, indicating that programs vary in their ability to improve classroom practice. Our results further indicate that

FIGURE 1
Number of References per Year of Randomized Field Studies That Target Classroom Practice



Note. The figure shows the citation returns from an EBSCO Host search of the previous 25 years, by year, for the following search string: ("classroom practice" OR instruction OR "instructional practice" OR "classroom practice" OR "teacher effectiveness") AND (intervention OR strateg* OR program OR treatment) AND (experiment OR "randomized experiment" OR "randomized trial" OR "randomized control").

interventions with a more limited dosage of treatment tend to produce similar effects to those with more intensive approaches. To present our study, the rest of the chapter is organized as follows. First, we summarize earlier studies that reviewed the research on teacher professional learning and present our theoretical framing. Next, we describe our approach. We then present our findings, and conclude with a discussion of the implications of our findings for practice and future research.

Previous Reviews of Research on Teacher Professional Learning

Several researchers have conducted reviews of research on the effects of interventions that target practice for teachers in K–12 settings. For example, Ingersoll and Strong (2011) focused on the effects of induction programs on beginning teachers and found positive effects for classroom practices in the majority of the 15 studies they reviewed. Slavin, Lake, Hanley, and Thurston (2012, 2014) focused on interventions of science instruction and found positive effects on student achievement for science teaching methods that focused on enhancing teachers' classroom instruction, but no effects for curriculum-focused teaching methods (i.e., programs that provide science kits to teachers), suggesting the importance of addressing classroom practice as a mediator to improving student outcomes. Gersten et al. (2009) found 42 studies to include in their syntheses of mathematics instructional interventions for students with disabilities and estimated positive and statistically significant mean effects for nearly all the aspects of classroom practice they studied. McKenna, Shin, and Ciullo (2015) also focused on instruction for students with disabilities and found some evidence of improved teacher use of targeted classroom practices across their 11 studies.

Other reviews have focused on sorting approaches used in teacher training and development, which may or may not contain an empirical synthesis. For example, Kennedy (2016) conducted a systematic review of the teacher professional development literature that focused on rigorous research studies that included student achievement outcomes. She identified 28 studies that met her inclusion criteria and was able to compute effect size estimates, which she sorted across the focal ideas teachers were expected to learn and the strategies for helping teachers execute those ideas in their practice. The review did not, however, contain a quantitative synthesis. Certainly many others have categorized the literature (e.g., Blank & de las Alas, 2009; Scher & O'Reilly, 2009; Timperley, Wilson, Barrar, & Fung, 2007). However, as Kennedy (2016) notes, most do not empirically examine the average effectiveness of these strategies for student or teacher outcomes, nor do they advance our understanding of differential effects among strategies. Most of these reviews also focus on understanding the relationship between professional learning strategies and student outcomes, without investigating the degree to which they affect intermediate outcomes like classroom practice.

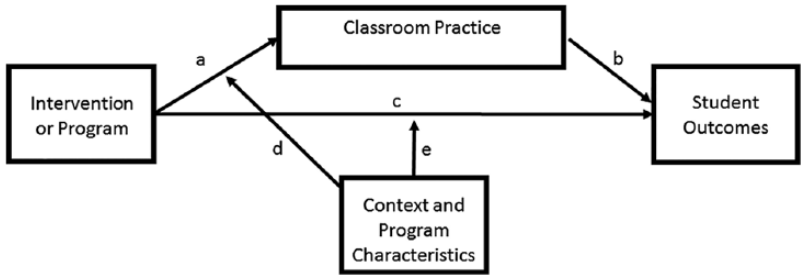
A more recent study by Kraft, Blazar, and Hogan (2018) conducted a meta-analysis of coaching intervention effects. The authors identified 60 studies for which the intervention included coaching for pre-K–12 teachers and included a measure of classroom practice or student achievement as an outcome. The authors found positive effects of coaching on both classroom practice (0.49 standard deviations) and student achievement (0.18 standard deviations). However, the study did not consider effects separately for teachers working in pre-K and K–12 settings, which is likely an area worth further understanding given the substantial differences between the pre-K and grade school teaching forces and their working conditions (Herzfeldt-Kamprath & Ullrich, 2016). Furthermore, while the study provides insights about the benefits of teacher coaching, more can be learned from the broader literature on teacher professional learning.

Thus, while these published reviews show positive effects for improving teachers' classroom practices and student achievement, they are limited in scope as each is focused on particular types of interventions (e.g., coaching, induction programs, or science instruction) and specific types of samples (e.g., beginning teachers or students with disabilities). More inclusive meta-analyses across a range of intervention types—for both pre-K and K–12 settings—can help the field better understand how professional learning programs can change teachers' classroom practices, and ultimately student achievement outcomes. In addition, further disaggregating the effects by the features of the intervention, sample, setting, and classroom observation measure is important to produce a clearer picture of what interventions work, for whom, and when.

Theoretical Framework

This work is guided by a simple theoretical framework that captures the underlying impetus behind the range of policies and resources dedicated to improving teacher

FIGURE 2
Theoretical Framework for Classroom Practice



effectiveness. As shown in Figure 2, an intervention or program directed toward classroom practice is hypothesized to bring changes in classroom practice (Path *a*) as well as changes in student outcomes (Path *c*). In this framing, changes in student outcomes have a direct relationship with the classroom practices they experience (Path *b*), and the impact of the intervention or program on student outcomes is mediated by the changes in classroom practice (Paths $a \times b$). Studies may examine only the direct effect of intervention on student outcomes (Path *c*) for a variety of design and logistical reasons and, therefore, may assume that observed effects on students happen indirectly, through Paths *a* and *b*, rather than through other programmatic mechanisms such as materials, curricula, or student supports. Also, there may be factors that moderate the programs' effects on classroom practice and student outcomes (Paths *d* and *e*), such as the context in which the program took place (e.g., characteristics of the students, teachers, or schools participating) or the features of the program (e.g., use of a coaching component, including a focus on data use).

This meta-analysis looks across the literature on classroom practice interventions to build knowledge about the extent to which this theoretical structure holds for Paths *a* and *d*. Existing quantitative syntheses that build knowledge across primary studies largely have focused on providing information for Paths *a* or *c*, but for specialized programs or populations. The research also offers little information to understand moderation effects that can account for the heterogeneity of program effectiveness across settings, populations, and interventions, illustrated by Paths *d* and *e*. This gap in the research represents a critical, untapped opportunity, given the more recent surge of rigorous research in classroom practice-directed interventions (see Figure 1).

APPROACH

Our study was designed to examine the responsiveness of classroom practice as measured through classroom observations. We define the study eligibility criteria, search strategy, study coding, effect size computation, and model estimation approach below.

Eligibility Criteria

To be eligible for inclusion in this meta-analysis, primary studies needed to meet the following criteria:

1. The study sample includes in-service teachers working in kindergarten to Grade 12.
2. The study evaluates an intervention that aims to improve classroom practices for supporting student academic learning (e.g., reading, math, science, social studies). Interventions may include professional development, training, and coaching for teachers. Curriculum interventions without a teacher training component that included a classroom practice focus were excluded.
3. The study design is a randomized control trial with randomization taking place at the teacher level or higher. The study must also include a control group.
4. The study uses a measure of classroom practice as measured through classroom observations.
5. The study is written in English.
6. The study provided sufficient information to calculate an effect size estimate and variance.

No restrictions were placed on actual study location or study year.

Literature Search and Retrieval

Our process of identifying relevant studies related to classroom practice effects as measured through classroom observations is graphically presented in Figure 3. We describe the process in detail below.

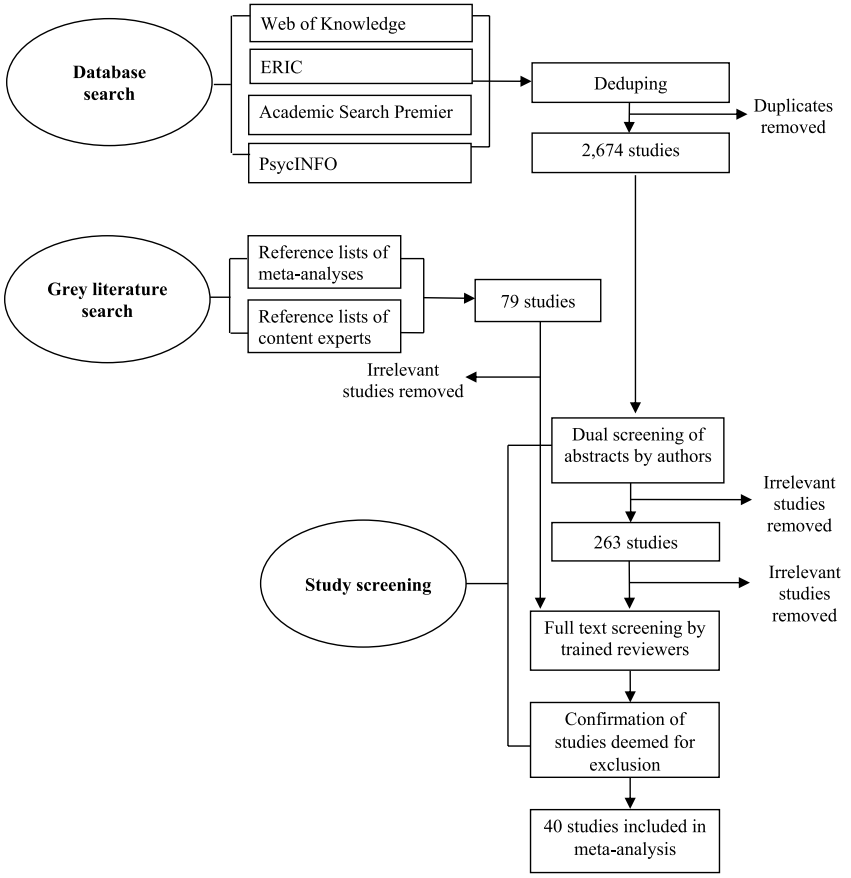
Database Search

We conducted electronic database searches of Web of Knowledge,[1] ERIC, Academic Search Premier, and PsycINFO on May 4, 2016. The search was limited to English language–only studies and studies including Grades pre-K–12.[2] No time frame restriction was imposed on the search.

We used the following search terms in our database search:

In title/subject/abstract: [Teacher OR educator OR "education* professional" OR instructor] AND ["Teach* practice" OR "instruction* approach" OR pedagogy OR practice OR enactment OR "classroom practice" OR "classroom performance" OR "teach* effectiveness" OR "teach* performance" OR "teach* efficacy" OR "instruction* efficacy" OR "teach* quality" OR "instruction* quality" OR "instruction* practice" OR "educat* practice" OR "educat* approach" OR "educat* quality" OR "educat* efficacy" OR "educat* effectiveness" OR "educat* performance"] AND [Intervention OR treatment OR program OR policy OR "professional development" OR training]
AND

FIGURE 3
Classroom Practice Meta-Analysis Literature Search and Retrieval Process



In full text: randomiz* OR "random* assign*"
NOT
In Subject: postsecondary or post-secondary or "higher ed*" or adult or nontraditional or college or universit*

After removal of duplicate studies, the database search yielded 2,674 studies.

Abstract and Title Screening

As represented in Figure 3, the three authors screened the titles and abstracts of studies using the eligibility criteria defined in the previous section. Because abstracts do not always specify whether the classroom practice measure is observation-based, we

only required that the abstract indicate that classroom practice was measured at all, rather than specifically measured using observations. All 2,674 studies were assigned to dual title and abstract screening, such that each study could be screened by two authors. Any discrepancies were resolved by consensuses between the two authors.

We conducted our title and abstract screening using abstrackr, a free, open-source tool that uses machine learning technology to semiautomate the screening process (Wallace, Small, Brodley, Lau, & Trikalinos, 2012). Using our criteria, we indicated whether a study was eligible for inclusion in our meta-analysis using the "Yes", "Maybe," and "No" options. Learning which studies are most pertinent, abstrackr prioritized the screening of studies most likely to be relevant to our meta-analysis. Of the 2,674 studies, we dual-screened 2,018 studies. At this point, abstrackr indicated that the remaining 656 studies had a less than 0.5 probability of inclusion (Wallace et al., 2012). The first author screened the titles (and, when relevant, abstracts) of the 656 low-probability studies to verify their exclusion and agreed that all 656 should be excluded. Of the 2,018 screened studies, 263 studies were labeled as "Yes" or "Maybe" for inclusion and the full text was pulled for further screening.

Gray Literature Search

We searched for gray, or unpublished, literature using two methods. First, we scanned the reference lists of 16 meta-analyses focused on examining classroom practice (identified using the search above). Second, we reached out to experts in the field who provided us with lists of studies that include interventions for teachers and observations of classroom practice. We screened the titles and abstracts of these studies using our eligibility criteria defined above. Excluding studies our previous search identified, this search yielded an additional 79 studies for potential inclusion and for which the full text was pulled for further screening.

Full Text Screening

Four screeners screened the full text of the 342 studies that were deemed for potential inclusion based on their titles and abstracts. Ten percent of the studies that were screened out by the screeners were also screened by one of the authors. All dual-screened studies marked for exclusion were confirmed for exclusion by one of the authors. We screened the full text using the following criteria: (1) the study is a randomized control trial; (2) the study includes an observation-based measure of classroom practice; (3) the study includes an analysis that compares treatment and control group teachers on the observation-based classroom practice measure; and (4) the study provides sufficient information to calculate an effect size estimate and variance. We also selected studies that included Grades K–12 samples and set aside studies solely focused on pre-K.[3] The interventions and outcomes found in the pre-K studies differ drastically from the types of interventions and outcomes in the K–12 studies; thus, we plan to conduct a separate meta-analysis focused on interventions for pre-K teachers. Forty studies met our final inclusion criteria.

Study Coding

Four coders coded the 40 studies included from our review. Twenty percent of the studies were also coded by one of the authors. Any coding discrepancies were examined and resolved by one of the authors.

We coded available information from the studies using Cronbach's (1982) UTOS (units, treatments, outcomes, and settings) framework for generalizability. Coding focused on extracting core descriptive information from the identified studies as well as characteristics that could explain, at least in part, observed heterogeneity in classroom practice effects (i.e., potential moderators). Before coding began, the research team identified items related to the content of this meta-analysis, such as intervention features, classroom practice domains, and measurement features, by drawing on the expertise of the research team and in consultation with other content experts who have conducted large-scale professional learning experiments. Our goal was to strike a balance between comprehensiveness and feasibility. As such, we recognize there may be a number of other study features that warrant further inquiry.

To code intervention features, the study team reviewed the qualitative descriptions provided in the texts. An intervention feature was coded as "present" if there was a clear description of the feature, and otherwise was considered not present if there was either (1) a clear description that the feature was not used or (2) there was no information to make a determination either way.

To code the classroom practice measures, the study reviewed the descriptions provided in each study. The study conceptualized broad domains and constructs within those domains using the observable portions of *The Framework for Teaching Evaluation Instrument* (Danielson, 2013) as a guideline. Classroom practices were categorized into either classroom environment or instructional domains where possible, and where the measures descriptions were too general to assign to either of those two domains, they were categorized into an overall effectiveness domain. Where feasible, we further categorized measures into constructs within the classroom environment and instructional domains. Within environment, we were able to identify measures that fell into categories including classroom culture, classroom management, or that aggregated over multiple aspects of environment. Within instruction, we were able to categorize measures that related to instructional format, discourse (which included questioning), student engagement and measures that again aggregated over multiple aspects of instruction. We also categorized practices that were specific to mathematics or English language arts content (e.g., use of decoding strategies).

A copy of the codebook is included in Appendix A (available in the online version of the journal). The codes included the following:

- *Study-level information*: authors, publication type, citation, and the year of publication (if published)
- *Sample characteristics (U)*: sample sizes (districts, schools, teachers, and students), student demographic composition (i.e., percent eligible for free or reduced price

lunch, percent sample minority, percent sample special education, and percent sample English learners), and years of teaching experience
- *Intervention characteristics (T)*: intervention features (e.g., coaching, video and technology components), delivery mechanisms (e.g., in person vs. online coaching), and intervention dosage (i.e., intervention length in weeks and hours spent on intervention)
- *Outcome information (O)*: the broad domains of the classroom practice outcomes measured in the study (e.g., classroom environment, instruction, overall effectiveness) and constructs within those domains (e.g., classroom management, instructional format), and observation timing (i.e., during intervention, directly after, after time passed, or a combination)
- *Setting information (S)*: grade level and level of randomization
- *Effect size information*: summary statistics of impact estimate (e.g., means and standard deviations, *t* tests, *F* tests, χ^2 tests, models, effect sizes and their types), and variance estimates of the outcome

Lack of information precluded the study team from coding for all of the information identified in the codebook. For example, the included studies often provided limited information on sample characteristics, and frequently did not include any information on the reliability of the classroom observation instruments used. The analyses therefore focused on the codes that produced sufficient information for empirical investigation.

Computing Effect Sizes

Meta-analysis relies on effect sizes, which provide a common metric for synthesis across studies that measure outcomes on different scales. Effect sizes encode both the direction and the magnitude of the relationship between intervention and outcomes (Hedges & Olkin, 1985; Lipsey & Wilson, 2001). For this meta-analysis, we computed the standardized mean difference (SMD) effect size for all classroom practice outcomes reported in each study. We computed SMDs using reported summary statistics, including means and standard deviations, *t* tests, *F* tests, χ^2 tests, regression model estimates, and effect sizes in other metrics. The equations for calculating the SMD, or converting other effect size metrics to the SMD, can be found in Borenstein, Hedges, Higgins, and Rothstein (2009).

Appropriate summary statistics were not available to calculate SMDs for all effects. We queried authors for the missing information, which yielded little extra information for coding purposes.

Hedges's Small Sample Correction

An SMD is generally estimated using Cohen's *d*-index:

$$d = \frac{\bar{y}^T - \bar{y}^C}{s},$$

where $\bar{y}^T$ and $\bar{y}^C$ are the sample means of the treatment and control groups, respectively, and s is the total pooled within-group standard deviation.[4] To account for small studies, we applied Hedges's (1981) small sample bias correction to the computed *d* effects:

$$g = d\left[1-\left(\frac{3}{4df-1}\right)\right],$$

where *df* denotes the degrees of freedom, equal to $N^T + N^C - 2$. The variance of *g* is

$$v_g = \frac{N^T + N^C}{N^T N^C} + \frac{g^2}{2\left(N^T + N^C - 2\right)},$$

where N^T and N^C are the total sample sizes of the treatment and control groups, respectively.

Adjusting for Nesting

In education, students are often nested within teachers who are then nested within schools and districts. In our set of studies, teachers were generally nested, or clustered, within schools. To account for this two-level nesting, when possible, we adjusted the computed effect sizes and their variances using Hedges's (2007) corrections[5]:

$$g_2 = g\sqrt{1-\frac{2*(n-1)*\rho}{N-2}}$$

and

$$V_{g2} = \left(\frac{N}{N^T N^C}\right)\left(1+(n-1)\rho\right)+\frac{g_2^2}{2h_{g2}},$$

where *N* is the total sample size in the study, *n* is the average number of teachers per school, ρ is the intraclass correlation coefficient,[6] and *h* is the effective degrees of freedom, given by

$$h_{g2} = \frac{(N-2)\left[(N-2)-2(n-1)\rho\right]}{(N-2)(1-\rho)^2 + n(N-2n)\rho^2 + 2(N-2n)\rho(1-\rho)}.$$

META-ANALYSIS

To estimate our meta-analytic mean effects, we employed random- and mixed-effects models for which studies are considered a random sample of possible studies, allowing us to generalize to a hypothetical population of studies by incorporating the between-study heterogeneity statistic τ^2 (Hedges & Vevea, 1998). The following is the mean-only random-effects model used to estimate the overall weighted mean of all observed classroom practice effects:

$$g_{jk} = \beta_0 + u_k + e_{jk},$$

where g_{jk} is the *j*th effect size estimate from study k[7]; β_0 is the mean effect; u_k is a study-level random error term, assumed to be normally distributed with a mean of zero, and between-study variance τ_u^2; e_{jk} is the sampling error for effect size j in study k, assumed to be normally distributed with a mean of zero and sampling variance σ_e^2; and the weights are computed as the inverse of the variance plus the estimated between-study variance, or $w_{jk} = 1/(v_{jk} + \hat{\tau}_u^2)$ (Borenstein et al., 2009). We employed the mean-only model to estimate weighted mean effects on *all* observed outcomes as well as separate effects for each classroom practice outcome domain and construct.

Examining Sources of Heterogeneity

To examine the heterogeneity of the observed classroom practice effects, we employed mixed-effects models that include intervention intensity (length and hours), timing of the observational measurement, intervention features, grade band, teacher experience, and study characteristics (publication year and sample size), separately, as moderators in the model. We tested moderators one at a time due to power constraints and also to avoid issues of correlation among the moderators (e.g., intervention length correlating with types of outcomes). For each of these moderators, the mixed-effects model is defined as follows:

$$g_{jk} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{jk} + \mathbf{Domain}_{jk} + u_k + e_{jk},$$

where β_1 is the estimated fixed effect of moderator X_{jk} and $\mathbf{Domain}_{jk}$ is a vector of fixed effects controlling for the classroom practice outcome domain.

Due to an excess of missing information, we were not able to estimate mixed-effects models for all of the variables coded (e.g., student characteristics). And, for the variables (or moderators) we did include in our analyses, only subsets of studies were included in each analysis due to some missingness across studies. As a result, we often lack power to detect statistically significant effects and discuss the results in terms of the magnitudes (or size) of the effects.

Estimating Heterogeneity

We estimated heterogeneity using I^2. I^2 represents the percentage of variation across studies that is due to heterogeneity rather than chance and may be thought of as the proportion of total variation in the treatment effects that is due to variation between studies (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003). The I^2 statistic may be estimated as follows:

$$I^2 = 100\left(Q - df\right)/Q,$$

where Q is Cochran's (1950) measure of heterogeneity:

$$Q = \Sigma w_{jk}\left(g_{jk} - \hat{\mu}_F\right)^2$$

and w_{jk} and $\hat{\mu}_F$ denote the weights and weighted average, respectively, from the fixed effects model.[8] Q follows a χ^2 distribution with $df = j - 1$.

To further characterize variation in how classroom practice responds to intervention, we also provide a 95% prediction interval for the estimated average effect (i.e., Borenstein, Higgins, Hedges, & Rothstein, 2017). The prediction interval provides the range of effect sizes that one would expect to see in future studies 95% of the time, based on the current analysis. Operationally, a 95% prediction interval is defined as follows:

$$PI_{95\%} = \hat{\beta}_0 \pm 1.96\left(\hat{\tau}\right),$$

where $\hat{\beta}$ is the average effect size estimate and $\hat{\tau}$ is the square root of the estimated between-study heterogeneity parameter (i.e., the estimated standard deviation of true population effects), estimated with restricted maximum likelihood.

Adjusting for Effect Size Dependencies

Most studies included multiple effects per study due to the measurement of multiple outcomes or samples. Effects from the same study are dependent on one another and may be correlated, and these dependencies are not fully corrected through specifying study-specific random effects in the analytic model. To account for these effect size dependencies and adjust the variance of the model estimates, we used robust variance estimation when estimating our random- and mixed-effects models. The equations for robust variance estimation can be found in Hedges, Tipton, and Johnson (2010) and Tipton (2014), and the R package, *robumeta*, is available in Fisher and Tipton (2015). We set ρ, the within-study effect-size correlation used to fit the correlated effects meta-regression models to 0.53, based on the overall mean effect across all studies.

Adjusting for Publication Bias

Although we tried to combat publication bias by searching for gray literature, we realize it is not possible to track down every unpublished classroom practice study ever conducted. Thus, we explored the impact of publication bias on each meta-analysis using Citkowicz and Vevea's (2017) beta-density weight-function model.[9] The selection model provides adjusted meta-analytic mean estimates and tests for publication bias, allowing us to examine the degree to which publication bias is an issue in the given meta-analysis. Moreover, the model allows for the inclusion of moderators into the model, allowing us to examine publication bias in the mixed-effects models.

FINDINGS

Results of the Search Process

Using this approach for the systematic literature search, we identified 40 studies for inclusion (see Figure 3 for a depiction of the literature search and retrieval process and Table 1 for the list of included studies). Because studies typically included multiple outcomes, we extracted 321 effects from the 40 included studies.

The studies were coded for all relevant effect sizes for impacts on instructional practice, in addition to coding for a range of intervention, observation measure and contextual factors. See Table 2 for summary information on the studies and the study features that were coded. Full references for the coded studies can be found in Appendix B (available in the online version of the journal).

Meta-Analytic Findings

In our discussion of the meta-analytic findings, we focus primarily on the magnitudes of the effects. Drawing on guidelines from the What Works Clearinghouse (Institute of Education Sciences, 2017), we consider effect sizes (or differences in effect sizes when making group comparisons) of 0.10 to 0.25 medium sized and suggestive, and effect sizes greater than 0.25 sufficiently large to be of substantive interest.

As explained earlier, the nature of moderator or subgroup meta-analyses typically offers limited power, and so relying solely on standard significance tests may be insufficient for interpreting results. Therefore, we examine significance and also discuss the effect sizes and the prediction intervals associated with the average effects.

We find that, on average, the randomized field trials targeting classroom practice yielded a positive, statistically significant effect of 0.42 (0.07) standard deviations based on classroom observations, as presented in Table 3. While this is promising, our results also give caution: We observed a substantial amount of heterogeneity across effect sizes. Our estimated I^2 indicates that 73% of the total variation in our estimates of the treatment effect is due to heterogeneity between effects rather than within effects (e.g., sampling error). Moreover, our absolute effect sizes range from –0.94 to

TABLE 1
Studies Included in Meta-Analysis

Study	Number of Effects	Number of Teachers	School Level	Outcome Domains	Intervention Features
Grigg, Kelly, Gamoran, and Borman (2013)	18	285	Elementary	Instruction	Group training
Cordray, Pion, Brandt, Molefe, and Toby (2012)	8	172	Elementary	Instruction	Individual training, group training, instructional materials
C. C. Johnson and Fargo (2010)	1	14	Middle	Overall effectiveness	Individual training, group training, instructional materials
Nelson-Walker et al. (2013)	13	42	Elementary	Instruction	Individual training, group training, instructional materials
Lowry (2007)	4	53	Elementary, middle, high	Instruction	Individual training, group training
Cappella et al. (2015)	4	120	Elementary	Classroom environment	Group training, instructional materials
Nugent et al. (2016)	6	92	Middle, high	Instruction, overall effectiveness	Individual training, group training
Gregory, Allen, Mikami, Hafen, and Pianta (2014)	5	87	Middle, high	Classroom environment, instruction	Individual training, group training
DeCesare, McClelland, and Randel (2017)	6	74	Elementary	Classroom environment, instruction, overall effectiveness	Individual training, group training
Parkinson, Salinger, Meakin, and Smith (2015)	2	130	Elementary	Classroom environment, instruction	Individual training, group training, instructional materials
Doabler (2010)	3	65	Elementary	Instruction	Individual training, group training
Reinke, Herman, and Dong (2014)	1	105	Elementary	Classroom environment	Individual training, group training

(*continued*)

TABLE 1 (CONTINUED)

Study	Number of Effects	Number of Teachers	School Level	Outcome Domains	Intervention Features
Wanzek et al. (2015)	2	24	Middle	Classroom environment, overall effectiveness	Individual training, group training, instructional materials
Abry, Rimm-Kaufman, Larsen, and Brewer (2013)	1	239	Elementary	Overall effectiveness	Individual training, group training, instructional materials
Meyers et al. (2016)	4	158	Middle	Classroom environment, instruction	Individual training, group training
Ottmar, Rimm-Kaufman, Larsen, and Berry (2015)	2	88	Elementary	Instruction, overall effectiveness	Individual training, group training
Evertson (1989)	69	29	Elementary, middle	Classroom environment, instruction	Group training
Goodson, Wolf, Bell, Turner, and Finney (2010)	3	128	Elementary	Classroom environment, instruction	Individual training, group training
Ottmar, Rimm-Kaufman, Berry, and Larsen (2013)	1	88	Elementary	Instruction	Individual training, group training
Simmons (2010)	5	60	Middle	Classroom environment	Individual training, group training
Motoca et al. (2014)	22	138	Middle	Classroom environment	Individual training, group training
Santagata, Kersting, Givvin, and Stigler (2011)	3	44	Middle	Instruction	Individual training, group training
Faraclas (2015)	6	24	Middle, high	Classroom environment, instruction, overall effectiveness	Individual training, group training
Supovitz (2013)	4	64	Elementary	Instruction	Individual training, group training
Doabler et al. (2014)	7	129	Elementary	Instruction	Group training, instructional materials
Jacob, Hill, and Corey (2017)	4	57	Elementary	Instruction	Group training
L. D. Johnson et al. (2014)	10	43	Elementary, middle	Classroom environment	Individual training, instructional materials

(*continued*)

TABLE 1 (CONTINUED)

Study	Number of Effects	Number of Teachers	School Level	Outcome Domains	Intervention Features
Garet et al. (2016)	6	165	Elementary	Instruction	Individual training, group training
Garet et al. (2008)	12	330	Elementary	Instruction	Individual training, group training
Brown, Jones, LaRusso, and Aber (2010)	3	82	Elementary	Classroom environment, instruction	Individual training, group training, instructional materials
Matsumura, Garnier, and Spybrook (2012)	3	93	Elementary	Instruction	Individual training, group training
Garet et al. (2017)	12	997	Elementary, middle	Classroom environment, instruction	Individual training, group training
Connor et al. (2011)	1	25	Elementary	Instruction	Individual training, group training, instructional materials
Vadasy, Sanders, and Logan Herrera (2015)	30	61	Elementary	Instruction	Group training, instructional materials
Gersten, Dimino, Jayanthi, Kim, and Santoro (2010)	6	81	Elementary	Instruction	Group training, instructional materials
Glazerman et al. (2008)	6	631	Elementary, middle	Classroom environment, instruction	Individual training, group training
Baker, Santoro, L., Biancarrosa, and Baker (2015)	1	39	Elementary	Instruction	Individual training, group training, instructional materials
Bos et al. (2012)	9	527	Middle	Instruction, overall effectiveness	Individual training, group training, instructional materials
Garet et al. (2010)	12	358	Middle	Instruction	Individual training, group training, instructional materials
Blazar and Kraft (2015); Kraft and Blazar (2013, 2017)	6	184	Elementary, middle, high	Classroom environment, instruction	Individual training, group training

TABLE 2
Study Characteristics

	N	*k*
Year of publication		
Prior to 2013	16	148
2013 or later	24	173
Teacher sample size		
Less than 100 teachers	28	222
100 or more teachers	12	99
Years of teaching experience		
Up to 10 years	14	91
More than 10 years	19	196
School levels included		
Grades K–5	29	246
Grades 6–8	17	182
Grades 9–12	5	27
Intervention features		
Individual training	33	177
Regular coach	10	59
Structured protocol for coaching	15	118
Instructional materials	16	110
Teacher-driven learning	6	26
Technology enhanced	9	57
Focus on using data to inform instruction	10	61
Active learning/practice in training	17	92
Intervention length		
≤26 weeks	8	99
>26 to ≤52 weeks	17	124
>52 weeks	9	67
Intervention hours		
≤20 hours	9	144
>20 hours to ≤100 hours	14	112
>100 hours	6	23
Timing of observational measurement		
Mid-stream/during	10	91
Directly after	30	151
After time passed	2	14
Combination	12	75
Total	40	321

Note. *N* = number of studies; *k* = number of effect sizes.

TABLE 3
Overall Mean Effects and Mean Effects by Instructional Practice Constructs

			Confidence Interval						Prediction Interval	
Construct	*M*	*SE*	Lower	Upper	*N*	*k*	I^2	τ^2	Lower	Upper
All effects	0.42	0.07	0.28	0.56	40	321	72.60	0.16	−0.38	1.22
Broad domain										
Classroom environment	0.27	0.08	0.10	0.44	17	112	51.76	0.07	−0.24	0.78
Instruction	0.46	0.08	0.29	0.63	32	198	76.69	0.18	−0.38	1.30
Overall effectiveness	0.49	0.29	−0.20	1.18	8	11	84.90	0.44	−0.82	1.79
Constructs within broad domains										
Classroom environment										
Aggregate environment[a]	0.32	0.13	−0.03	0.67	6	22	52.37	0.07	−0.21	0.85
Classroom culture	0.16	0.06	0.01	0.31	9	32	41.02	0.03	−0.21	0.53
Classroom management	0.43	0.24	−0.17	1.02	7	58	57.02	0.24	−0.52	1.38
Instruction										
ELA content	0.55	0.16	0.13	0.96	6	49	84.70	0.30	−0.53	1.62
Math content	0.31	0.04	0.20	0.41	7	21	0.00	0.00	0.31	0.31
Instructional format	0.21	0.08	0.03	0.40	10	22	16.07	0.01	0.00	0.43
Discourse[b]	0.28	0.10	0.05	0.50	11	43	53.07	0.08	−0.26	0.82
Student engagement	0.46	0.12	0.14	0.79	6	11	69.12	0.11	−0.18	1.10
General/aggregate instruction[c]	0.64	0.20	0.20	1.07	15	35	85.78	0.34	−0.50	1.78

Note. Rho (ρ) was set to 0.53, based on the overall mean effect across all studies. ELA = English language arts; *N* = number of studies; *k* = number of effect sizes; I^2 = proportion of the total variance due to between effect size heterogeneity; τ^2 = total effect size heterogeneity.

[a]"Aggregate environment" includes measures that aggregated across two or more items capturing classroom environment.

[b]"Discourse" construct includes questioning, discussion, and formative assessment.

[c]"General/aggregate instruction" includes measures that aggregated across two or more items capturing instruction, or overall instructional measures.

2.76 standard deviations, with a 95% prediction interval of –0.38 to 1.22. This indicates that while classroom practice is responsive to intervention, there is substantial variation in the average population effects. Additionally, in our subgroup analyses that focus on specific aspects of instruction, the findings consistently demonstrate positive and mostly significant results across the different constructs, ranging from 0.16 (0.06) for classroom culture to 0.64 (0.20) for general instruction (see Table 3). Again, the findings further indicate substantial variability in effects across studies (with the exception of mathematics content) as observed by the wide confidence and prediction intervals and the high I^2s, ranging from 16% to 86%.

Additionally, we considered how the timing of a classroom observation measurement could be associated with the effect sizes, presented in Table 4. We did this through a moderator analysis that considered differences among effects that were measured during an intervention (i.e., while the intervention was ongoing and active), immediately after it ended (i.e., the first postintervention outcome assessment), or after some period of time had passed after the intervention (i.e., after the first postintervention outcome assessment). We found the largest effects among classroom observations measured right after an intervention completed (0.31), and only slightly smaller effects after time had passed (0.29) and during the intervention (0.26). While the differences were not statistically significant, this suggests that teachers may experience contemporaneous improvements in practice during interventions, which may taper off some shortly after an intervention ends but then plateau and potentially maintain for some time.

Intervention Intensity

We conducted additional moderator analyses to explore differences in effect sizes by the length of time over which an intervention occurred, the number of hours of the intervention, and when the observation of instruction occurred (Table 4). Appendix D (available in the online version of the journal) presents supplemental moderator analysis tables that include confidence and prediction intervals for the conditional means.

We first explored how intervention intensity could relate to effects by considering the total weeks the intervention lasted from start to end. Based on the patterns we observed in our data, we created three mutually exclusive categories, indicating whether an intervention lasted 26 weeks or fewer, over 26 and up to 52 weeks, or over 52 weeks long. We used the category indicating the shortest duration (26 weeks or fewer) as our reference group and conducted moderator analyses comparing the two longer duration categories to this group. We found the largest mean effects among interventions that lasted 26 or fewer weeks (0.43) compared with those that lasted 26 to 52 weeks (a mean of 0.28, or –0.15 *SD* lower than the reference group), and compared with interventions that lasted over 52 weeks (a mean of 0.24, or –0.19 *SD* lower than the reference group). Again, while the differences were not significant, we consider the magnitude of the differences suggestive. While not shown, when plotting effects by length of intervention, we noticed that studies of interventions with a

TABLE 4
Mean Effects by Timing of Observations and Intervention Intensity

Intervention Feature	*M*	*SE*	*N*	Δ From Reference Group	*SE*	*p* Value
Timing of observational measurement						
Mid-stream/during (reference group)	0.26	0.13	10			
Directly after	0.31	0.10	30	0.05	0.12	0.67
After time passed	0.29	0.50	2	0.03	0.51	0.96
Combination	0.36	0.15	12	0.10	0.17	0.56
Intervention length						
≤26 weeks (reference group)	0.43	0.23	8			
>26 to ≤52 weeks	0.28	0.17	17	−0.15	0.31	0.63
>52 weeks	0.24	0.16	9	−0.19	0.30	0.54
Intervention hours						
≤20 hours (reference group)	0.37	0.14	9			
>20 hours to ≤100 hours	0.39	0.14	14	0.02	0.16	0.90
>100 hours	0.33	0.16	6	−0.04	0.19	0.83

Note. Rho (ρ) was set to 0.53, based on the overall mean effect across all studies. *N* = number of studies; Δ = difference between the reference group and group of interest (i.e., the slope estimate $\hat{\beta}_g$).

longer duration had more variation in effects (i.e., the effects were not as consistent) compared with shorter duration studies, which tended to cluster around a particular mean effect.

To approach the question of intervention intensity slightly differently, we also examined differences in means by the number of program hours reported. Again, we empirically derived three categories based on our data, this time categorizing interventions according to whether they lasted 20 hours or fewer, over 20 up to 100 hours, or over 100 hours. We used the smallest category (20 hours or fewer) as our reference group to compare against the two longer duration categories. We found no significant or substantive differences comparing those with 20 or fewer hours to interventions with 20 to 100 hours or greater than 100 hours. This suggests that teachers are just as likely to benefit in less intensive interventions than more intensive ones.

Intervention Features

Of the intervention features that we analyzed in the moderator meta-analyses (Table 5), we did not find any statistically significant outcomes indicating particularly salient approaches to professional learning, but the magnitude of the results suggests a number of potential insights.

We examined effect sizes associated with interventions that had an individualized training component. Of the 33 interventions that included an individualized

TABLE 5
Mean Effects by Intervention Features

Intervention Feature	*M*	*SE*	*N*	Comparison Feature	*M*	*SE*	*N*	Δ	*SE*	*p* Value
Individual training	0.33	0.09	33	No individual training	0.17	0.11	8	0.16	0.11	0.17
Established coach	0.14	0.13	10	Other coach	0.42	0.12	18	0.28	0.16	0.09
In-person + remote training	0.43	0.22	6	In-person only	0.31	0.13	23	0.12	0.25	0.64
Structured observation protocol for coaching	0.21	0.10	15	No structured observation protocol used/described	0.49	0.15	18	−0.28	0.17	0.11
Active learning/practice	0.43	0.15	17	No active learning/practice described	0.25	0.09	24	0.18	0.15	0.23
Focus on using data to inform instruction	0.45	0.22	10	No focus on using data	0.26	0.10	30	0.19	0.24	0.44
Instructional materials	0.38	0.14	16	No instructional materials	0.27	0.10	24	0.11	0.17	0.53
Teacher-driven learning	0.34	0.25	6	Not teacher driven	0.31	0.09	34	0.03	0.26	0.92
Technology enhanced	0.37	0.19	9	No technology	0.29	0.11	31	0.07	0.23	0.75
Summer + school year	0.29	0.09	21	No summer	0.39	0.12	19	−0.10	0.08	0.23

Note. Rho (ρ) was set to 0.53, based on the overall mean effect across all studies. *N* = number of studies; Δ = difference between the group means (i.e., the slope estimate $\hat{\beta}_1$).

component (most often coaching), the average effect was 0.33, which was 0.16 standard deviations higher than the mean effects from the 8 interventions that did not include an individualized component (with a mean of 0.17). In a closer look among the subset of 33 interventions with individualized training, we compared effect sizes for interventions that used an established coach or trainer, such as using an established building or district-based coach or a full-time professional development provider, compared with effect sizes where someone took on the coach or trainer role specifically for the purposes of the study. We found that mean effects for the 10 studies that had an established coach (0.14) were smaller than the mean effects for the 18 studies that had a coach who either took on the coach role specifically for study or for whom there was no description (0.42). The difference in effects was 0.28 standard deviations, although this was again not statistically significant. We also found that interventions using a combination of remote and in-person coaching had higher mean effects by 0.12 standard deviations compared with interventions with only in-person training. Surprisingly, we found lower mean effects (0.28 standard deviation lower average) among studies that identified using a structured protocol for observation and feedback (0.21), compared with studies that allowed ad hoc feedback processes or that simply didn't specify the process in the intervention description (0.49).

We found several other potential indications of useful intervention features, all with insignificant but medium-sized differences in mean effects comparing among groups. Specifically, we found positive differences in group means in favor of interventions that provide teacher active learning opportunities to apply and practice the instructional skills during the training (0.18), training teachers to use data to guide their instruction (0.19), and including instructional materials (0.11). On the other hand, interventions that lasted over the school year and summer (0.29) produced smaller effects than interventions that did not include time over the summer (0.39). The other intervention features we considered—descriptions of teacher-driven learning and technology-enhanced learning experiences—yielded small changes, falling below our threshold of 0.1 to indicate something of potential substantive interest.

Grade Band and Teacher Experience

When considering the grade span of teachers included in a study (Table 6), our findings suggest that, compared with studies that included elementary school teachers (0.30), studies with middle school teachers had smaller average effects (0.27) and studies with high school teachers had larger mean effects (0.49). Neither of these were significantly different. We view the differences by grade band as suggestive of differences worth noting. However, given the small number of studies and effect sizes that include high school teachers, we interpret these results with extra caution.

TABLE 6
Mean Effects by Grade Band and Teacher Experience

	No			Yes					
Teaching Grade Span	*M*	*SE*	*N*	*M*	*SE*	*N*	Δ	*SE*	*p* Value
Grades K–5	0.33	0.17	11	0.30	0.10	29	−0.03	0.19	0.86
Grades 6–8	0.38	0.11	23	0.27	0.11	17	−0.11	0.16	0.49
Grades 9–12	0.28	0.09	35	0.49	0.33	5	0.20	0.35	0.59
More than 10 years of teaching experience	0.12	0.11	14	0.36	0.13	19	0.24	0.16	0.14

Note. Rho (ρ) was set to 0.53, based on the overall mean effect across all studies. *N* = number of studies; Δ = difference between the group means (i.e., the slope estimate $\hat{\beta}_1$). Studies with teachers across multiple grade bands are included in each category; therefore, the groups are not mutually exclusive.

While information on years of experience for teacher samples was often unreported or reported inconsistently, we were able to classify samples as averaging more than 10 years of experience or less among the studies that presented teacher experience information. Analyses indicated that impacts based on samples of teachers with 10 or more years of experience were higher (0.24 standard deviations) than impacts from samples of teachers with fewer than 10 years of experience, although for both groups the means were still positive (0.36 for the more experienced samples and 0.12 for the less experienced samples). Unfortunately, due to limited and inconsistent reporting about teacher characteristics across studies, we were not able to examine other teacher characteristics.

Study Features

Table 7 presents results from several additional analyses that consider features of the research studies. First, we wanted to consider whether the field was potentially getting more effective in developing useful teacher learning experiences, and so we looked separately at studies published more recently, that is, 2013 or later, compared with those published before 2013. We did not find meaningful differences in this comparison. Given the recent interest in coaching in particular, we looked at studies using this publication year divide among the interventions that included an individualized training component. We found some suggestion that more recent studies have been more effective (0.17 *SD* higher means) among these studies.

Finally, we compared impacts from studies with fewer than 100 teachers with those with 100 or more teachers. Analyses indicate an average effect size that is 0.25 standard deviations larger among the studies with fewer teachers, which meets our criteria for being of substantive interest, and this is also the one contrast we found that was statistically significant. While not examined through this study, one potential explanation is that studies with more teachers represent scale-up studies and

TABLE 7
Mean Effects by Study Features

	No			Yes					
Study Feature	*M*	*SE*	*N*	*M*	*SE*	*N*	Δ	*SE*	*p* Value
Study published in 2013 or later	0.25	0.13	16	0.33	0.10	24	0.08	0.15	0.61
Study published in 2013 or later (individual component only)	0.19	0.14	14	0.36	0.11	19	0.17	0.17	0.32
Total teacher *N* was less than 100	0.16	0.11	12	0.41	0.10	28	0.25	0.12	0.05

Note. Rho (ρ) was set to 0.53, based on the overall mean effect across all studies. *N* = number of studies; Δ = difference between the group means (i.e., the slope estimate $\hat{\beta}_1$).

reflect the difficulties of scaling teacher professional development. Alternatively, this may be a sign of publication bias, since studies with fewer teachers would need larger effects to be statistically significant, and therefore of interest for publication. We discuss publication bias next.

Publication Bias

The results of the publication bias analyses indicate that substantial publication bias is present in our set of meta-analyses. Using Citkowicz and Vevea's (2017) beta-density weight-function model, the overall estimate of 0.34[10] was adjusted upward to 0.86 standard deviations.[11] This result implies that larger effect size estimates with smaller *p* values are underrepresented in the literature that we synthesized; the opposite direction of what is typically associated with publication bias. The adjustment might not be as substantial with the inclusion of quasi-experimental study effect size estimates, given that they tend to be larger than randomized controlled trial results (Cheung & Slavin, 2016). Tables C1 and C2 in Appendix C (available in the online version of the journal) also present the publication bias results for the individual construct analyses and moderator analyses. With a few exceptions, the results are similar across all analyses such that the adjusted estimate is larger when adjusted for publication bias. While the true population effects are likely somewhere between the unadjusted and adjusted estimates, our analyses do not provide strong evidence of selective reporting as a function of statistical significance. That is, the estimated average effect is likely an underestimate of the true population effect.

DISCUSSION

Is Classroom Practice Responsive When Targeted Through Professional Learning Programs?

Our results indicate that interventions directly targeting classroom practice through professional learning can bring about meaningful shifts in

practice as measured through classroom observation. We find that the effects of these interventions are on average positive, and the magnitude of the changes in classroom practice are substantively notable. Moreover, these findings generalize across multiple domains and subconstructs of observed classroom practice. This indicates that the responsiveness of classroom practice is not tied to a specific aspect of observable practice, and likewise that interventions can support teachers in developing different types of classroom practice skills. From this, we conclude that teachers stand to benefit from professional learning opportunities designed to promote their classroom practice development across a range of areas.

Moreover, our results suggest that teachers can continue to develop their expertise over their career. Although we could only examine the relationship of teacher experience with responsiveness of practice in a coarsened way due to limitations in reporting of teacher experience in the studies, our results are suggestive of greater intervention effects for samples of teachers with an average teaching tenure of more than 10 years. This may be surprising given prior research noting that novice teachers have lower instructional effectiveness and make particularly large improvements during their first years on the job (Desimone, Hochberg, & McMaken, 2016; Kini & Podolsky, 2016). Given the poor quality of information available for us to meta-analyze teacher experience, we do not wish to make too much of this finding, other than to point out the implication that classroom practice likely is a malleable factor for teachers across a range of tenure years, and even experienced teachers may be able to make observable and substantive shifts in their practice.

Can We Identify "Active Ingredients" in Professional Learning That Are Associated With Positive Changes in Classroom Practice?

While our findings provide strong evidence for the responsiveness of classroom practice when targeted through professional learning opportunities, the findings also highlight an important caution about the wide variability of how much classroom practice actually changes through a given intervention or implementation. The studies in this meta-analysis yielded effect sizes that varied from large and negative, to small and negligible, and to large and positive. This variation creates uncertainty about how much classroom practice will change in a given instance, and requires further exploration to understand identifiable sources of the heterogeneity in classroom practice outcomes.

One key question is then whether we can identify specific approaches to professional learning that appear particularly beneficial to promoting positive shifts in classroom practice; our findings did not uncover any "silver bullet" to promote effective classroom practices, but the results do suggest several things. First, our findings indicate the promise of including some kind of individual, personalized training component for teachers to develop their classroom practice. This likely aligns with conventional wisdom that individualized professional learning (e.g., coaching) is beneficial—reflected in the fact that only 8 of the 41 interventions[12] studied did not

include some kind of individual component—but our meta-analysis provides empirical evidence to support the idea.

Beyond individualized training, our results suggest two other professional learning features worth noting. Our findings suggest the benefit of allowing teachers active learning opportunities where they may directly engage with or apply what they are learning during the context of the training. This coincides with earlier research that has identified the importance of allowing teachers active learning opportunities in professional development for yielding improvements in instruction (Desimone, Porter, Garet, Yoon, & Birman, 2002) and research that has found classroom practice improves when teachers are given opportunities for deliberate practice of instruction across multiple settings (Ericsson, 2006; Grossman, Hammerness, & McDonald, 2009; Lampert et al., 2013). We used a liberal approach to capturing active learning opportunities, considering it part of an intervention either through a specific description of the active learning opportunity or if the study simply stated that the approach included active learning, for example. We recognize that we were only able to capture this very broadly, which likely implies measurement error, but that measurement error would attenuate the findings and so the extent to which we see positive benefits in light of this is particularly interesting. Last, we also find suggestive evidence in support of helping teachers use student data to inform their instruction. Research has pointed to the increased emphasis of data-driven decision making to inform instruction over the last decade in response to accountability policy initiatives, but actual practices implemented vary and are inconsistent (Datnow & Hubbard, 2015). With the increased prevalence of available data and the suggested benefit through these findings, this may be an aspect of professional learning worth further study.

In contrast to these areas of promise, our findings also shed some light on what may be less useful or necessary for supporting development of teacher classroom practice. Specifically, we find no indication that interventions that have higher number of hours or last over a longer window of time provide enhanced benefits to classroom practice. In some ways this is counterintuitive, since *prima facie*, further support would presumably be linked to increased benefits. However, this result was also found in a meta-analysis of coaching studies (Kraft et al., 2018). One potential explanation is that longer, more intensive professional learning interventions may include a broader focus, while shorter interventions may include more narrowly targeted focal areas, producing larger effects in those targeted areas. For example, some researchers argue that teachers master complex forms of instruction when they first work on smaller routines and then build up to more complex activity (e.g., see Cai et al., 2017a, 2017b; Kraft & Blazar, 2017).

Is Classroom Practice Responsive Enough to Target in Short-Cycle Professional Learning Approaches?

Building on the finding that shorter, less intensive interventions yielded similar outcomes to those that were longer and more intensive, our results further suggest

that classroom practice is malleable and responsive to intervention quickly enough to target in short-cycle professional learning approaches. We find that classroom practice effects measured midstream to the intervention were not meaningfully lower compared with when they were observed after the intervention, indicating that it is responsive and improvements can be found even before an intervention is fully implemented. Taken together with the findings that a larger dosage is not necessarily needed, this suggests that short-cycle, continuous improvement efforts to address classroom practice may be successful, and in fact may be an efficient way to support improvements in classroom practice.

What Are the Implications for Scaling Interventions to Improve Classroom Practice?

While these findings support the promise of improving classroom practice through intervention, they may also suggest the challenges in scaling programs. As discussed previously, our analysis contrasting studies with greater than 100 teachers to those with fewer indicated smaller benefits to classroom practice among studies including more teachers. This was the only moderator we analyzed that was both of substantive size and statistically significant. We also found that among studies with an individualized training component, there were larger effects in studies that either described using a coach that provided services specifically for the study or did not provide a description of the coach background compared with studies that described coaches as individuals who regularly provide coaching outside the study context. This may possibly reflect the difficulties of scale-up as well, since delivering professional learning across a larger group of teachers typically requires working with local coaches and trainers. In our data, the average number of teachers per study with an established coach was 154.5 teachers, compared with 111.6 teachers among studies that engaged people to coach specifically for study purposes or did not describe the coaches' backgrounds, lending some support to this connection. In smaller studies, such as efficacy trials, the program developers may be more likely to deliver the training and require fewer coaches. Larger studies may require more coaches and introduce greater variation in coaching effectiveness and implementation fidelity. Taken together, our findings provide further impetus to the field that continued research and work is needed to understand how to deliver effective professional learning opportunities focused on classroom practice at scale.

Directions for Future Research

Educator instructional effectiveness continues to be an important area of policy and a focal area for educators where substantial resources are spent. The findings from this meta-analysis provide strong empirical support for the potential to improve classroom practice through professional learning interventions, although research must continue to seek the most effective ways to provide useful learning opportunities. One critical next step will be to continue the meta-analysis of these studies to identify links

between changes in classroom practice with changes in student outcomes. While large-scale, rigorous research on content-focused professional development has produced disappointing results in terms of improvements to student achievement, there is a need to identify the professional learning approaches or types of classroom practices that are most often associated with improved student outcomes and thus may provide a focus for new professional learning programs (Garet, Heppen, Walters, Smith, & Yang, 2016), and further meta-analyses tying changes in classroom practice to changes in student outcomes may help meet this need.

Our results also provide guidance for the design of research studies. For example, the average effect sizes identified through this work can inform power calculations when designing new research. These results support powering a study for larger effect sizes for classroom practice outcomes, compared with effects for student achievement outcomes (Hedges & Hedberg, 2007). The heterogeneity of the distribution of observed effects can also inform Bayesian analytic approaches that need estimates of not only the mean but also the level of uncertainty based on the expected distribution. Furthermore, our ability to understand effects of interventions would be enhanced by more in-depth information about implementation and implementation context (e.g., Hill, Corey, & Jacob, 2018). This could be achieved through mixed-methods approaches that provide deeper and qualitative explorations of how interventions were implemented and teacher experiences both during the professional learning and as they seek to apply their learning to the classroom. Future research thus can use the insights from these findings to inform evaluation design and analysis, as we work as a field to continue to build knowledge in the critical area of effective teaching.

ACKNOWLEDGMENTS

The authors thank Jane Coggshall, Mike Garet, Josh Polanin, Andrew Wayne, and members of the O2 Lab at American Institutes for Research for helpful comments and suggestions.

NOTES

Supplemental material is available for this article in the online version of the journal.

[1]The Web of Knowledge search was performed only in Social Science Citations and Social Science Conference Proceedings. Moreover, Web of Knowledge does not support full-text searching, thus the search was conducted in the default fields, as opposed to the fields specified in our search terms.

[2]The studies were grouped into those focused solely on Pre-K and Grades K–12 due to their differing focuses on interventions and outcomes only after the literature search was completed.

[3]The authors are currently in the process of conducting a separate study focusing on pre-K samples.

[4]The pooled standard deviation may be computed by $\sqrt{\frac{\left(N^T-1\right)S_T^2+\left(N^C-1\right)S_C^2}{N^T+N^C-2}}$.

[5]Adjustment is not necessary if teachers are randomly assigned within schools and school fixed effects are included in the analysis model.

[6]The intraclass correlation represents the amount of variance that is found between the

clusters (or teachers), relative to the total variation. It may be estimated by $\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$, where σ_B^2 and σ_W^2 are the between- and within-cluster variances. These were rarely reported by study authors. For studies where ρ, or information to calculate ρ, was not reported, we imputed the average of all calculated ρs, 0.18.

[7]When each sample in the analysis produces a single effect size, $j = k$.

[8]The fixed-effects model is estimated using the same approach as the random-effects model, with the exception that the fixed-effects model assumes that the effects are homogenous and omits the between-study heterogeneity statistic τ^2 from the model.

[9]Citkowicz and Vevea's (2017) beta-density weight-function model uses the beta density to explicitly model the selection process, or process by which studies are assumed to be chosen for publication. It then adjusts the meta-analytic results for publication bias by multiplying the usual probability density function for the effect-size model (i.e., a standard meta-analytic model) by the selection model.

[10]The overall estimate presented in Table 3 (0.42) is computed while adjusting for effect size dependencies. Currently, no publication bias method exists that allows for the simultaneous adjustment of effect size dependencies and publication bias. Thus, our publication bias adjustment is conducted on the meta-analytic mean that is unadjusted for effect size dependencies (0.34).

[11]We also used Vevea and Hedges's (1995) step-function model to adjust the meta-analytic estimates for publication bias and obtained similar results. The overall estimate was also adjusted upward, to 0.68 standard deviations.

[12]One study included two interventions, resulting in 41 interventions across 40 studies.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*, 95–135.

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. Retrieved from https://eric.ed.gov/?id=ED544700

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis* (1st ed.). Chichester, England: Wiley.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*, 5–18. doi:10.1002/jrsm.1230

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., & Hiebert, J. (2017a). Clarifying the impact of educational research on learning opportunities. *Journal for Research in Mathematics Education*, *48*, 230–236.

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., & Hiebert, J. (2017b). Making classroom implementation an integral part of research. *Journal for Research in Mathematics Education*, *48*, 342–347.

Chamberlin, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(43), 17176–17182.

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*, 283–292. doi:10.3102/0013189X16656615

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods*, *22*, 28–41. doi:10.1037/met0000119

Cochran, W. G. (1950). The comparison of percentage in matched samples. *Biometrica*, *37*, 256–266. doi:10.1093/biomet/37.3-4.256

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs* (1st ed.). San Francisco, CA: Jossey-Bass.

Danielson, C. (2013). *The framework for teaching evaluation instrument*. Princeton, NJ: Danielson Group.

Datnow, A., & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers College Record*, *117*(4), 1–26.

Desimone, L., Hochberg, D., & McMaken, J. (2016). Teacher knowledge and instructional quality of beginning teachers: Growth and linkages. *Teachers College Record*, *118*(5), 1–54.

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

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–703). Cambridge, England: Cambridge University Press.

Fisher, Z., & Tipton, E. (2015). *robumeta: Robust variance meta-regression* (R package version 1.6). Retrieved from http://CRAN.R-project.org/package=robumeta

Garet, M. S., Heppen, J. B., Walters, K., Smith, T. M., & Yang, R. (2016). *Does content-focused teacher professional development work? Findings from three Institute of Education Sciences studies* (NCEE 2017-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20174010/pdf/20174010.pdf

Gersten, R., Chard, D., Jayanthi, M., Baker, S., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*, 1202–1242.

Goldhaber, D. D. (2002). The mystery of good teaching. *Education Next*, *2*, 50–55.

Grossman, P., Hammerness, K., & McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching: Theory and Practice*, *15*, 273–289.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. doi:10.2307/1164588

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370. doi:10.3102/1076998606298043

Hedges, L. V., & Hedberg, E. (2007). Interclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Herzfeldt-Kamprath, R., & Ullrich, R. (2016). *Examining teacher effectiveness between preschool and third grade*. Washington, DC: Center for American Progress. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2016/01/19064517/P-3TeacherEffectiveness2.pdf

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558. doi:10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistencies in meta-analyses. *British Medical Journal*, *327*, 557–560.

Hill, H., Corey, D., & Jacob, R. (2018). Dividing by zero: Exploring null results in a mathematics professional development program. *Teachers College Record*, *120*(6), 1–42.

Ingersoll, R. M., & Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers, the impact of induction and mentoring programs for beginning teachers: A critical review of the research, a critical review of the research. *Review of Educational Research*, *81*, 201–233. doi:10.3102/0034654311403323

Institute of Education Sciences. (2017). *What Works Clearinghouse: Standards handbook* (Version 4.0). Washington, DC: U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Jacob, A., & McGovern, K. (2015). *The mirage: Confronting the hard truth about our quest for teacher development*. Retrieved from https://tntp.org/assets/documents/TNTP-Mirage_2015.pdf

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

Kini, T., & Podolsky, A. (2016). *Does teaching experience increase teacher effectiveness? A review of the research*. Palo Alto, CA: Learning Policy Institute, 2016. Retrieved from https://learningpolicyinstitute.org/sites/default/files/product-files/Teaching_Experience_Report_June_2016.pdf

Kraft, M. A., & Blazar, D. (2017). Individualized coaching to improve classroom practice across grades and subjects: New experimental evidence. *Educational Policy*, *31*, 1033–1068.

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*, 547–588. doi:10.3102/0034654318759268

Lampert, M., Franke, M., Kazemi, E., Ghousseini, H., Turrou, A. C., . . . Crowe, K. (2013). Keeping it complex: Using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education*, *64*, 226–243.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

McKenna, J., Shin, M., & Ciullo, S. (2015). Evaluating reading and mathematics instruction for students with learning disabilities: A synthesis of observation research. *Learning Disability Quarterly*, *38*, 195–207.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica*, *73*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*, 247–252.

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.

Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2012). *Effective programs for elementary science: A best-evidence synthesis*. Baltimore, MD: Johns Hopkins University, Center for Research and Reform in Education.

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*, 870–901.

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration*. Wellington, New Zealand: Ministry of Education.

Tipton, E. (2014). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*, 375–393. doi:10.1037/met0000011

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*, 419–435.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 819–824). New York, NY: Association for Computing Machinery.