The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence

Matthew A. Kraft
Brown University

David Blazar
Harvard University

Dylan Hogan
Brown University

November 2016

Updated: June 2017

Abstract

Teacher coaching has emerged as a promising alternative to traditional models of professional development. We review the empirical literature on teacher coaching and conduct meta-analyses to estimate the mean effect of coaching on teachers' instructional practice and students' academic achievement. Combining results across 44 studies that employ causal research designs, we find pooled effect sizes of .58 standard deviations (SD) on instruction and .15 SD on achievement. Much of this evidence comes from literacy coaching programs for pre-kindergarten and elementary school teachers. Although these findings affirm the potential of coaching as a development tool, further analyses illustrate the challenges of taking coaching programs to scale while maintaining effectiveness. Coaching effects in large-scale effectiveness trials with 100 teachers or more are only half as large as effects in small-scale efficacy trials. We conclude by discussing ways to address scale-up implementation challenges and providing guidance for future causal studies.

Suggested Citation:
Kraft, M.A., Blazar, D., Hogan, D. (2016). The effect of teaching coaching on instruction and achievement: A meta-analysis of the causal evidence. Brown University Working Paper.

Check here for the most up-to-date version

Correspondence regrading the article can be send to Matthew Kraft at mkraft@brown.edu. We thank Robin Jacob, Sara Rimm-Kaufman, Kiel McQueen, Robert Pianta, and Beth Tipton for their feedback at various stages the research and the many authors who responded to our queries. Adam Merier provided excellent research assistance. All mistakes all our own.

The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence

Providing high-quality professional development to employees is among the most important and longstanding challenges faced by organizations. Investments in on-the-job training offer large potential returns to workforce productivity. However, high-quality programs have proven difficult to develop, scale, and sustain. These challenges are particularly acute in the public education sector given the size of the teacher labor market and the dynamic nature of the job. Every day, over 3.5 million teachers in the United States (U.S.) face unique challenges educating students who enter the classroom with a wide range of knowledge, skills, and needs.

Across the U.S., school systems spend tens of billions of dollars annually on professional development (PD) to help teachers meet these daily challenges with limited results to show for these investments.[1] Impact evaluations find that PD programs more often than not fail to produce systematic improvements in instructional practice or student achievement, especially when implemented at-scale (Jacob & Lefgren, 2004; Garet et al., 2008; Garet et al., 2011; Garet et al., 2016; Glazerman et al., 2010; Harris & Sass, 2011; Randel et al., 2011). These findings are particularly troubling given the wide variation in effectiveness across teachers and the lasting impact teachers have on long-term student outcomes in the labor market and beyond (Chetty, Friedman, & Rockoff, 2014; Jackson, 2016). Both of these findings make improving the skills of the teacher workforce a societal and economic imperative (Hanushek, 2011). The need for further training has only grown in recent years as professional expectations for teachers continue

[1] Arriving at an exact estimate of total expenditures on PD is complicated by the fact that federal requirements have districts report expenditures on PD as part of an "Instructional staff services" category which also includes expenditures for curriculum development, libraries, and media and computer centers. Most studies find that districts allocate 3% to 5% of their total budget to support teacher development (Odden, Archibald, Fermanich, & Gallagher, 2002; Miles, Odden, Fermanich, Archibald, & Gallagher, 2004). Given that total expenditures for U.S. K-12 public schools were $620 billion in 2012-13, even a conservative estimate puts this number in the tens of billions (Jacob & McGovern, 2015).

to rise and states adopt new “college- and career-ready” standards that require teachers to integrate higher-order thinking and social-emotional learning into the curriculum.

The failure of traditional PD programing to improve instruction and achievement has generated calls for research to identify specific conditions under which PD programs might produce more favorable outcomes (Desimone, 2009; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). These efforts have led to a growing consensus that effective PD programs share several “critical features” including job-embedded practice, intense and sustained durations, a focus on discrete skill sets, and active-learning (Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009 ; Desimone, 2009; Desimone & Garet, 2015; Garet, Porter, Desimone, Birman, & Yoon, 2001; Hill, 2007). A recent meta-analysis found that math- or science-oriented PD programs with many of these features were associated with improvements in both instructional practices and academic achievement (Scher & O’Reilly, 2009). However, this review identified only one randomized control trial, and many of the quasi-experiments it included “had significant methodological weaknesses” (p.223). Kennedy’s (2016) findings from a graphical analysis of popular design features in PD programs were more mixed: a focus on content knowledge, collective participation, or intensity did not appear to be associated with program effectiveness. We extend this work by reviewing the causal evidence on one specific PD model that is centered on several of these “critical features” and that has gained increasing attention in recent years: teacher coaching.

Teacher coaching has a deep history in educational practice. Pioneering work by Joyce and Showers in the 1980’s helped to build the theory and practice of teacher coaching as well as some of the first empirical evidence of its promise (Joyce & Showers, 1982; Showers, 1984, 1985). They conceptualized coaching as an essential feature of PD training that facilitates teachers' ability to translate knowledge and skills into actual classroom practice (Joyce & Showers, 2002). The practice of teacher coaching remained limited in the 1980's and 1990's with most programs developing out of local initiatives. Beginning in the late 1990's, federal legislation aimed at strengthening the quality of reading instruction helped formalize and fund coach positions for reading teachers in schools (Denton & Hasbrouck, 2009). These included the passage of the Reading Excellence Act in 1999, No Child Left Behind (NCLB) in 2002, and the reauthorization of the Individuals with Disabilities Education Act (IDEA) in 2004. The legacy of these investments is evident today in the wide range of established literacy coaching programs and the preponderance of research focused on literacy coaching models.

Existing handbooks and reviews of the teacher coaching literature have focused on describing the theory of action, creating typologies of different coaching models, and cataloguing best implementation practices (Cornett & Knight, 2009; Devine, Meyers & Houssemand, 2013; Fletcher & Mullen, 2012; Kretlow & Bartholomew, 2010; Obara, 2010; Schachter, 2015; Stormont, Reinke, Newcomer, Marchese, & Lewis, 2015). Responding to the call by Hill, Beisiegel, and Jacob (2013) in their proposal for new directions in research on teacher PD, we complement these works by conducting the first meta-analysis of studies examining the causal effect of teacher coaching on instructional practice and student achievement.

This work would not have been possible only a decade ago. In 2007, a comprehensive review of the entire canon of teacher development literature found that only nine out of over 1,300 studies were capable of supporting causal inferences (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). The passage of the Education Sciences Reform Act (ESRA) in 2002, which authorized the Institute for Education Research (IES), raised the standards for methodological rigor in educational research and created new funding sources for large-scale program evaluation studies. IES-funded grants, combined with a growing movement calling for the wider adoption of causal inference methods in educational research (Cook, 2001; Angrist, 2004; Murnane & Nelson, 2007; Wayne et al., 2008), served to catalyze a new wave of randomized trials evaluating coaching and other PD programs.

Our review of the literature identified 44 studies of teacher coaching programs in the U.S. that used both a causal research design and examined effects on instruction or student achievement.[2] The use of meta-analytic methods to analyze these studies affords the ability to answer several macro- and micro-level questions about teacher coaching that no single experimental trial can address. First, we are able to better understand the efficacy of coaching as a general class of PD by analyzing results across a range of coaching models. Second, the large financial and logistical costs of conducting experimental evaluations of teacher coaching programs has resulted in many individual studies that are underpowered. Meta-analysis techniques leverage the increased statistical power afforded by pooling results across multiple studies. This is critical for determining whether common findings of positive effect sizes that are not statistically significant are due to limited statistical precision or chance sampling differences. Third, meta-analytic regression methods facilitate a comparison of different coaching models and a closer examination of specific design features that may drive program effects, such as the size of coaching programs, pairing coaching with other PD elements, in-person versus virtual coaching, or coaching dosage (Blazar & Kraft, 2015; Marsh et al., 2008; Ramey et al., 2011).

Our analyses are driven by three primary research questions:

RQ1: What is the causal effect of teacher coaching programs on classroom instruction and student achievement?

RQ2: Are specific coaching program design elements associated with larger effects?

[2] Studies included in the meta-analysis are marked with an "*" in the references.

> RQ3: What is the relationship between coaching program effects on classroom instruction and student achievement?

We pair empirical evidence from these analyses with a discussion of the implementation challenges and potential opportunities for scaling up high-quality coaching programs in cost effective ways. We then conclude with recommendations on how future studies can strengthen and extend the existing body of causal research on teacher coaching. By examining these questions, we hope to shed light on the efficacy of teacher coaching as a model of PD and inform ongoing efforts to improve the design, implementation, and studies of coaching programs.

Methods

Working Definition of Teacher Coaching Interventions

Although the majority of teacher coaching models share several key program features, no one set of features defines all coaching models. At its core, "coaching is characterized by an observation and feedback cycle in an ongoing instructional or clinical situation" (Joyce & Showers, 1981, p.170). Coaches are thought to be experts in their field who model research-based practices and work with teachers to incorporate these practices into their own classrooms (Sailors & Shanklin, 2010). However, in our review of the literature we encountered multiple, sometimes conflicting, working definitions of teacher coaching. Some envision coaching as a form of implementation support to ensure that new teaching practices – often taught in an initial training session – are executed with fidelity (Devine et al., 2013; Kretlow & Bartholomew, 2010). Others see coaching as a direct development tool that enables teachers to see "how and why certain strategies will make a difference for their students" (Russo, 2004, p. 1; see also Richard, 2003). Still others describe multiple types of coaching, each with their own objectives. For example, "responsive" coaching aims at helping teachers reflect on their practice, while “directive” coaching is oriented around the direct feedback coaches provide to strengthen teachers’ instructional practices (Ippolito, 2010). In line with these multiple perspectives, Gallucci et al. (2010) describe coaching as “inherently multifaceted and ambiguous” (p. 922). Coaches often take on these roles and others, including identifying appropriate interventions for teacher learning, gathering data in classrooms, and leading whole-school reform efforts.

To arrive at a working definition of coaching, we situate it within a broader theory of action around teacher PD, which we outline in Figure 1. The ultimate goal of teacher PD program is to support student learning and development broadly defined but often operationalized narrowly as performance on standardized achievement tests (Devine et al., 2013; Desimone, 2009; Kennedy, 2016; Schachter, 2015). Mapping backwards, many argue that student achievement will not increase without changes in teacher knowledge or classroom practice (Cohen & Hill, 2000; Kennedy, 2016; Scher & O’Reilly, 2009). Training sessions, which are a standard form of PD offered to teachers (Darling-Hammond et al., 2009; Hill, 2007), are thought to be beneficial in improving teachers’ knowledge. However, this approach often is viewed as insufficient to address the inherently multifaceted nature of teachers’ practice and how they enact their knowledge and skills in the classroom (Kennedy, 2016; Opfer & Pedder, 2011; Schachter, 2015). Teacher coaching is considered a key lever for improving teachers’ classroom instruction and for translating knowledge into new classroom practices. To do so, coaches engage in a sustained “professional dialogue” with coachees focused on developing specific skills to enhance their teaching (Lofthouse, Leat, Towler, Hall, & Cummings, 2010).

Because improvements in teacher skill and classroom practice cannot be divorced from improvements in teacher knowledge (Hill, Blazar, & Lynch, 2015), coaching rarely is implemented on its own. Often, coaching is combined with training sessions or courses in which teachers are taught new skills or content knowledge (Kretlow & Bartholomew, 2010). It also may be used to develop teachers’ abilities to work with new curricular materials or instructional resources. In a review of the literature on PD in early childhood settings, Schachter (2015) found that 39 of the 42 programs that included coaching as one element combined it with some other form of training (e.g., a workshop or course), and many also included additional resources such as curriculum materials or websites with video libraries.

We define coaching programs broadly as all in-service PD programs that incorporate coaching as a key feature of the model. The role of the coach may be performed by a range of personnel including administrators, master teachers, curriculum designers, and external experts. We characterize the coaching process as one where instructional experts work with teachers to discuss classroom practice in a way that is (a) *individualized* – coaching sessions are one-on-one; (b) *intensive* – coaches and teachers interact at least every couple of weeks; (c) *sustained* – teachers receive coaching over an extended period of time; (d) *context-specific* – teachers are coached on their practices within the context of their own classroom; and (e) *focused* – coaches work with teachers to engage in deliberate practice of specific skills. This definition is consistent with the research literature and allows us to include a wide spectrum of models in this analysis that range from those focused on supporting the implementation of curriculum or pedagogical frameworks to those where the coaching process itself is the core development tool.

For the purposes of this review, we narrow this definition in several ways that we see as consistent with the broader literature on coaching programs. First, we exclude teacher preparation and school-based teacher induction programs. While these types of teacher training are increasingly integrating observation and feedback cycles with instructional experts into their designs, it is difficult to disentangle coaching practices from the range of supports provided to new teachers as part of comprehensive induction programs (e.g., Glazerman et al., 2010). The role and goals of a mentor are often quite distinct from those of a coach. Second, we exclude programs in which teachers' classroom colleagues serve in a coach-like role. We recognize that peer-to-peer feedback has been a longstanding practice in the field (see, for example, Showers, 1985 for theory, and Papay, Taylor, Tyler & Laski, 2016 for a recent evaluation). However, we see the peer-to-peer dynamic as distinct from the expert role that coaches take on in the studies we review. Similarly, we exclude studies that employed non-experts such as research assistants (e.g., Cabell et al., 2011). Finally, we exclude coaching programs where coaches provide direct service to students in addition to supporting teachers (e.g., Raver et al., 2009), given that the pathway to improved student performance may work outside of instructional improvement.

Literature Search Procedures

We conducted a systematic review of the research literature through a three-phase process. We first identified articles using the electronic databases Academic Search Premier, Econ Lit, Ed Abstracts, ERIC, Google Scholar, ProQuest, and PsycINFO. We searched databases using the primary terms "*teach** AND *coach**" or "*professional development*" and then refined searches by combining these with the following terms: "*in-service*", "*model**", "*evaluation*", "*effect**", "*impact**", "*random**", "**experiment**", and "*trial*." Second, we reviewed references in prior reviews of coaching programs identified above and iteratively checked the references from the studies that met our inclusion criteria to cross-check the search process. Finally, we contacted leading scholars in the field including many authors of the articles included in this analysis to solicit their help in identifying additional causal analyses of teacher coaching.

Inclusion Criteria

We restricted the sample of studies published during or before 2016 using four primary criteria pertaining to the sample, the intervention, the research design, and the outcomes[3]. First, we required that studies evaluate a PD program that incorporated teacher coaching as defined by our working definition above. Second, we limited this review to include studies where the sample was comprised of early childhood to 12th grade teachers in the U.S.[4] Third, we required that studies employed an experimental or quasi-experimental research design capable of supporting causal inferences (Shadish, Cook, & Campbell, 2002; Murnane & Willett, 2011). We judged quasi-experimental designs as meeting this standard if they employed a regression discontinuity (no qualifying studies found), an instrumental variables approach with a justifiable instrument (no qualifying studies found), or a difference-in-differences design (e.g., Teemant, 2014; Vogt & Rogalla, 2009; Biancarosa, Bryk, & Dexter, 2010; Lockwood, McCombs, & Marsh, 2010). We excluded studies that relied principally on covariate adjustment or used a pre-post design for treated units only given concerns that these strategies cannot adequately account for non-random selection. Fourth, we required that studies include at least one measure of a teacher's classroom instruction as rated by an outside observer, or a measure of student achievement from a standardized assessment. We focused narrowly on these two classes of measures as they are directly aligned with the intended effect of coaching in our theory of change model. They also are the only two types of outcomes that were used regularly in most studies.[5] As causal research on teacher coaching continues to accumulate, meta-analytic work may

[3] When multiple papers were published using the same set of data, we included papers when they reported results from different outcomes (Rimm-Kaufman et al., 2014 and Abry, Rimm-Kaufman, Larsen, & Brewer, 2013), different cohorts (Kraft & Blazar, in press and Blazar & Kraft, 2015), or different years (Matsumura, Garnier, & Spybrook, 2013 and Matsumara, Garnier, & Spybrook, 2012) but chose only one of the studies when the samples, outcomes and periods of measurement were overlapping (Vernon-Feagans, Kainz, Hedrick, Ginsberg & Amendum, 2013instead of Amendum, Vernon-Feagans, & Ginsberg, 2011).

[4] For example, we excluded international studies such as Bowne, Yoshikawa, & Snow (2016), Rezzonico et al. (2015), Sailors et al. (2014), and Yoshikawa et al. (2015).

[5] Other types of outcomes included measures of teachers' core content knowledge, measures of teachers' content knowledge for teaching, and a range of social-emotional outcomes from student self-reports and teacher surveys. While these outcomes are of real importance, they were collected in very few studies.

examine effects on other outcomes, including teacher knowledge, to examine whether the entire theory of action presented in Figure 1 is borne out in the data. In the next section, we describe additional constraints placed on how these outcome measures were captured.

Outcomes

Instruction. Following the conceptual framework developed by Cohen, Raudenbush, and Ball (2003), we viewed instruction not simply as how teachers deliver lessons but rather as the interaction of teachers, students, and content within the context of classroom and school environments. Thus, we included scores from classroom observation instruments that capture teachers' pedagogical practices (e.g., the use of open-ended questions), as well as measures of teacher-student interactions (e.g., relationships), student-content interactions (e.g., student engagement), and the interactions among teachers, students, and content (e.g., classroom climate). We limited measures of instruction to include only those that were collected by outside observers blind to treatment status.[6] We excluded any measures that were self-reported by teachers to protect against self-report or reference bias.

Although a growing body of research drawing on data from observation instruments identify several unique domains of teaching practice (Blazar, Braslow, Charalambous, & Hill, 2017; Hamre et al., 2013), it was not feasible to examine these constructs separately in these analyses. Studies used several different observation instruments or coding schemes that aimed to capture different elements of teachers' instructional practice; these instruments tended to align with the goals of the specific coaching program or the grade level of the students in the classroom. These instruments included observation rubrics that are well-established in the research literature and widely used by districts (e.g., Classroom Assessment Scoring System

[6] The number of observations per teacher varies considerably across studies. We do not impose a minimum number of observations per teacher as an inclusion criteria.

[CLASS], Early Language and Literacy Classroom Observation [ELLCO]), as well as lesser-known instruments that were developed by the researchers or coaching program under study (e.g., Blazar & Kraft, 2015; Sailors & Price, 2015; Teemant, 2014). Because studies provided varying levels of information about these instruments, we were limited in our ability to assess the degree of overlap among specific dimensions. Relatedly, without access to the primary data, it was not possible to assess the measurement properties of scores produced by each of these instruments. However, most studies either used validated scales (e.g., CLASS, ELLCO), or reported strong reliability indices (e.g., 80% or higher inter-rater agreement rates, internal consistency reliability of 0.80 or higher).

Student achievement. We included in these analyses impacts on students' performance from a range of standardized achievement tests. These included both low-stakes and high-stakes standardized assessments administered as part of the normal schooling process as well as those administered specifically for research purposes. The vast majority of these measures were widely used assessments with well-established psychometric properties. Low-stakes assessments included the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), the Group Reading Assessment and Diagnostic Evaluation (GRADE), and the Peabody Picture Vocabulary Test (PPVT). High-stakes assessments were typically from mandatory end-of-year state tests such as the Virginia Standards of Learning (SOL) assessments and the Texas Assessment of Knowledge and Skills (TAKS). Several studies also administered assessments constructed using existing test-items from the Northwest Evaluation Association and The Trends in International Mathematics and Science Study (TIMSS). We view all of these assessments as aiming to capture student learning broadly. When feasible, we disaggregate results by subject.

Coding Procedures

We coded studies for information needed to convert treatment effects on instruction and achievement to Cohen's *d* (standardized effect sizes) and associated standard errors. We also developed codes for a range of study characteristics and coaching model features through an iterative process informed by theory, past meta-analytic studies, and patterns that emerged during our review of the literature. Each study was coded by at least two of the authors. Instead of conducting duplicate blind coding of each study, we sought to minimize error through a process of critical review (Dietrichson, Bøg, Filges & Jørgensen, 2017; Jacob & Parkinson, 2015). One author coded a study and a second author read the study and reviewed the codes to assess their accuracy. When discrepancies arose, all three authors conferred and worked to arrive at a consensus decision. We describe the codes used to characterize study features below:

Source and year of publication. We categorized the source of studies into two codes: those published in peer-reviewed journals or institute reports. Institute reports include contract research reports submitted to the federal government and studies conducted by large-scale contract research firms such as Mathematica Policy Research and RAND.

Research design. We organized studies into two categories: randomized control trials and quasi-experimental methods.

Level of randomization. We coded the level at which the researchers randomized entities into treatment and control conditions. These included randomization at the teacher, school, and district level.

Teacher sample size. We coded studies for the number of teachers included in the largest analytic sample as a proxy measure for the size of a coaching program.

School level. We created a set of four indicators for the level of schooling that was the focus of each study. These codes included pre-Kindergarten, Elementary (Kindergarten – 5th grade), Middle (6th – 8th grade), and High School (9th – 12th grade). Studies were coded in more than one category when they included teachers from grades that spanned multiple categories.

Coaching model type. We developed a set of codes for categorizing coaching models that was informed by existing theory and practical considerations for defining classifications to be broad enough to include a sufficient number of studies for meta-analytic purposes. We first divided the sample into studies of coaching that were focused on general pedagogical practices (e.g., programs that focused on improving students' social and emotional skills, including their behavior in class) versus those that were content-specific. We created these codes to be mutually exclusive, such that any study that included some focus on content-specific coaching was coded as such. Next, we coded content-specific studies into subgroups based on the specific subject areas that they addressed (i.e., reading, mathematics, science).

Complementary treatment elements. Many of the studies included in the sample combined teacher coaching with additional features of PD programming. We categorized these additional features into three broad codes: Group Trainings, capturing any workshops or trainings that teachers attended in addition to receiving one-on-one coaching; Instructional Content, capturing resources that teachers received (e.g., curriculum materials) that complemented their work with a coach or where the coach was meant to help the teacher implement these resources in the classroom; and Video Libraries, capturing instances in which teachers were provided with access to video recordings of other teachers' classroom instruction that served a core function in teachers' conversations with their coach. Through an interactive process, we found that these three codes captured nearly all additional and complementary resources that teachers received.

Mode of delivery. We coded coaching models as either delivered in person or virtually through web-based platforms. In one instance where coaching was delivered as a combination of both we coded the model as in-person coaching (Powell, Diamond, Burchinal, & Koehler, 2010) given that a one-time in-person meeting may be central to establishing productive relationships.

Coaching and total PD dosage. To the extent possible, we coded the average number of hours teachers worked one-on-one with a coach. We view this measure as exploratory given two measurement concerns. Sufficient information to calculate an estimate of coaching dosage was not always reported. Even when data was reported, studies sometimes differed in their characterization of the number of hours spent with a coach. In some instances, this included the total number of hours spent meeting with a coach either in-person or virtually. In other instances, authors included time coaches spent observing teachers as part of their description of coaching dosage. Where possible, our measure of coaching dosage excludes time spent in other PD activities such as summer workshops. We included this code in our analyses despite some reservations about its reliability in order to further explore the widely cited implications from Yoon et al.'s (2007) review that PD must be high dosage in order to be effective.

In many instances, coaching programs were paired with other PD features. To capture the full scope of the PD teachers received, we also coded the total number of reported hours that all elements of the PD program entailed. This, of course, cannot account for differing number of hours spent using support materials such as video libraries.

Teacher and Coach Characteristics. We also searched articles for information about teacher and coach characteristics but found that inconsistent reporting approaches and a lack of details limited our ability to construct formal codes. For example, authors most often reported information on teachers' years of teaching experience, but varied widely on how they reported this information (e.g., mean and standard deviation, percentages of teachers who fell into discrete experience bins, range). For coach characteristics, authors were even less consistent in what they reported. Some provided information on teaching experience, while others focused on the training provided to coaches.

Meta-Analytic Approach

We arrive at pooled effect sizes using meta-analytic methods that produce precision weighted estimates and account for the clustered nature of the data (Hedges, Tipton, & Johnson, 2010; Tanner-Smith, Tipton, & Polanin, 2016). Our inclusion criteria and coding process produced a total of 155 effect sizes for instructional outcomes and 82 effect sizes for achievement outcomes across the 44 studies. Many studies contributed more than one effect size for a given outcome type because multiple measures were used (e.g., studies that reported dimension-level scores from an observation instrument of teachers' classroom practice), or because measures of the same type were captured at multiple points in time. Some studies also included multiple effect sizes due to multiple treatment groups (e.g., PD workshop, coaching plus PD workshop, and business-as-usual control in Garet et al., 2008). Here, we focused only on the treatment-control contrast that most closely matched the designs of other studies: coaching (plus any complementary activities) versus business-as-usual control.

We estimate a standard random effects meta-analytic model where effect-sizes are viewed as data sampled from a distribution of true effects produced by a spectrum of coaching program models as follows:

$$y_{ij}^{k} = \alpha + u_j + \varepsilon_{ij}^{k} \qquad (1)$$

Here, y_{ij}^{k} captures a given effect size *i* for outcome type *k* in study *j* where models for different outcome types are fit separately. Alpha, α, captures the pooled effect size estimate for outcome k, u_j is the study level random effect, and ε_{ij}^{k} is the mean-zero stochastic error term.

We examine the association between components of different coaching models and effect-size outcomes by expanding this model to fit a meta-analytic regression as follows:

$$y_{ij}^{k} = \alpha + \beta' X_j + u_j + \varepsilon_{ij}^{k} \tag{2}$$

where X is a vector of study characteristics and β captures the estimates relating these characteristics and our outcomes of interest.

We estimate all models using Robust Variance Estimation (RVE) methods (Hedges et al., 2010; Tanner-Smith et al., 2016) which account for both the differing degrees of precision across studies as well as the non-independence of effect sizes within studies through a method that is analogous to clustered standard errors. Weights are constructed such that:

$$w_{ij}^{k} = \frac{1}{n_j^{k}(v_{.j} + \tau^2)} \tag{3}$$

where $v_{.j}$ is the mean of the individual *i* variances for the n_j effect sizes in study *j* for outcome *k*, and τ^2 is the estimated between-study random effect variance component from equation (1) [i.e., $Var(u_j) = \tau^2$] estimated via methods of moments. As equation 3 shows, effect sizes that are estimated with greater precision (due to differences in sample sizes, level of randomization, predictive power of covariates, etc.) are given larger weights; effect sizes from studies that contribute multiple effect size estimates are given less weight.

Results

Characteristics of Included Studies

Our search yielded a total of 44 studies that met the inclusion criteria. We present descriptive statistics on these studies in Table 1 and include the full list of studies and associated codes in Appendix Table A1. Thirty-two studies included observation ratings of teachers' instruction, 23 studies examined achievement outcomes, and 11 studies captured both outcomes. Every study we identified was published on or after 2006 with the vast majority of studies in peer-reviewed journals (n=38). Forty of the 44 studies employed experimental research designs. Twenty-nine studies evaluated content-specific coaching programs while 15 assessed coaching programs for general instructional pedagogy. Given the history of federal investments in literacy coaches, it should not be surprising that nearly all of the content-specific coaching models focused on reading and literacy (n=25 for reading, compared to n=2 for math and n=2 for science). Thirty-six of the 44 studies included teachers who worked in pre-kindergarten centers or elementary schools, another consequence of the early support for literacy coaching programs. Nine of the studies evaluated virtual coaching models where teachers recorded themselves teaching and discussed their instruction on a web-based platform with a virtual coach. Of these nine virtual coaching studies, seven evaluated versions of the My Teaching Partner program developed by Robert Pianta and colleagues at the University of Virginia Center for Advanced Study of Teaching and Learning.

Across the studies we examined, 91% evaluated coaching models that were combined with at least one additional PD element. This finding is nearly identical with Schachter's (2015) review of the literature on PD for pre-kindergarten educators. Coaching was combined most frequently with group trainings in the form of summer workshops and team training sessions during the academic year where coaches might demonstrate lessons or instructional practices (37 of 44). Eighteen of the 44 studies also provided teachers with instructional content materials such as curriculum, lesson plans, or guide books. Another eight studies supplemented coaching with video exemplars of other teachers delivering high-quality instruction.

We found that the reported number of hours teachers worked one-on-one with a coach varied widely across coaching programs. Nine studies reported coaching dosages of ten hours or less while twelve studies reported 21 hours or more. The total PD hours for participating teachers also varied widely across programs with eight interventions consisting of 20 total hours or less and nine interventions consisting of 60 total hours or more. This wide variation in the dosage of coaching and total PD hours illustrates the substantial differences in the coaching programs included in this meta-analysis.

Because average teaching experience was not reported in a consistent metric across studies, we do not include this information in Table 1. For those studies that did report mean years of teaching experience, the average was approximately 11 years. Some studies focused specifically on early career teachers (e.g. Blazar & Kraft, 2015), while others focused on more veteran teachers (e.g. Pianta, Mashburn, Downer, Hamre, & Justice, 2008; Teemant, 2014; Vernon-Feagans et al., 2013; Vogt & Rogalla, 2009).

Effects on Instruction and Achievement

Kernel density plots of effect sizes on teachers' instruction and students' achievement help provide visual evidence and intuition for our pooled estimates. As shown in Figure 2, the distribution of effect sizes of coaching on instruction is distributed approximately normally with a long right-hand side tail. The magnitude of effects vary considerably, with an interquartile range between .18 SD and .96 SD. Effects on achievement also are distributed approximately normally with a positive skew and an interquartile range between .04 SD and .22 SD.

Turning to our primary meta-analytic results for instruction in Table 2, Column 1, we find large positive effects of coaching on teachers' instructional practice. We find a pooled effect size of .58 standard deviations (SD) across all 32 studies that included a measure of instructional practice as an outcome. The associated standard deviation of the estimated random effect (⍰), a measure of the variation in effect sizes across programs, is .36 SD suggesting there exists substantial variability across programs. Disaggregating these results among content-specific coaching programs and those that focused on general pedagogical practices produces strikingly consistent estimates of .58 SD and .63 SD, respectively. The content-specific coaching programs covered several different areas: reading, mathematics, and science. However, only studies in reading had sufficient sample sizes to report disaggregated results, which also are quite similar. Of the two math-specific coaching programs and the two science-specific coaching programs, only one each included instruction as an outcome measure.

Teacher coaching also has a positive effect on student achievement as shown in Table 2, Columns 2-5. Across all coaching models, we estimate that coaching raised student performance on standardized tests by .15 SD based on effect sizes reported in 23 studies that included measures of students' academic performance. The associated estimate for ⍰ is .15 SD, again suggesting effects differ substantially across programs. The overall effect size estimate pools achievement tests across reading, math, and science in order to provide a broad picture of coaching effectiveness. However, our ability to generalize across subjects is limited by the fact that almost three quarters of the total number of achievement effect sizes are for reading. Pooled effects on science achievement from two studies (both of which evaluate science-focused coaching programs) are 0.11 SD, while effects on math achievement are smaller at .02 across four studies; neither estimate is statistically significant. In a supplemental analysis where we focus only on math-specific coaching program, we find that the estimate for math achievement increases to .08 (*p*=.44, k=14, n=2). The largest effects we find on achievement are from coaching programs targeting teachers' instruction around reading skills, which have an average effect of .18 SD on students' reading achievement. Effects for reading coaching programs are likely larger, in part, because of their focus on early childhood and elementary education where students make the largest learning gains (Lipsey et al., 2012).

We also see slightly smaller effects on student achievement for general coaching programs (.10 SD) than content-specific programs (.16 SD). This makes sense given that general coaching programs often are focused less on helping teachers improve students' test scores and more on their ability to engage students around their social and emotional development. This is also evident in the fact that only three of the eleven studies that evaluated general coaching programs examined effects on student achievement. However, due to small sample sizes for achievement effects of general coaching programs, we cannot statistically distinguish these estimates from each other.

Next, we explore potential differences in coaching program effects across school levels by estimating effects for pre-kindergarten centers, elementary school, middle school and high school separately. As shown in Table 3, the pattern of results is suggestive of slightly larger effects on instruction in pre-kindergarten and elementary schools. The pooled effect size for instruction is .66 SD and .58 SD for pre-kindergarten educators and elementary school teachers compared to .44 SD and .47 SD for middle school and high school teachers. For achievement outcomes, pooled effects are again larger in pre-kindergarten (.18 SD) and elementary schools (.17 SD) compared to middle schools (.09 SD). Achievement effects in high school are twice as large (.18 SD) as those in middle school but are limited to data from only three studies.

Features of Effective Coaching Programs

Coaching models differ both in their focus and their program features. We conduct exploratory analyses to examine whether certain program features are associated with larger or smaller pooled effect sizes. We emphasize that, despite the fact that we restrict the analytic sample to studies that employ causal research designs, these meta-analytic regressions do not capture the causal effect of a given program feature. Variation in these coaching features across programs is not random.

While these analyses are motivated by key questions in the research literature on the design of coaching models, we recognize two important limitations on statistical power that prevent us from ruling out smaller relationships in many cases. First, features of coaching models vary at the study level rather than effect-size level. Second, power for meta-analytic regressions is reduced by the unbalanced distribution of many of the predictors in the data (Tanner-Smith et al., 2016). Given these challenges, we report results in Appendix Tables A2 and A3. Here, we do not find any clear evidence of systematic differences in effect sizes based on features of the coaching model. This includes differences in instruction and achievement when coaching is combined with additional PD features, or when it is delivered in person versus virtually.

One exception is the exploratory analysis for dosage. For both measures of dosage – total hours of coaching, and total hours of PD when coaching is paired with other program features – we find relatively precise estimates of zero for instruction and achievement outcomes. Further, we do not find any clear evidence of potential threshold effects or other non-linear functional forms when we model these relationships using a set of four indicators. These findings are consistent with Kennedy's (2016) graphical analysis of features of effective PD programs, but stand in contrast to previous findings on the importance of dosage in PD programs more broadly (Yoon et al., 2007). These findings suggest that the quality and focus of coaching may be more important than the actual number of contact hours.

Does Better Instruction Lead to Higher Achievement?

A fundamental assumption underlying the theory of action for coaching and many other development models is that helping teachers improve the quality of their instructional practice will lead to improvements in student achievement (Cohen & Hill, 2000; Kennedy, 2016; Scher & O'Reilly, 2009; Weiss & Miller, 2006). Our coded meta-analysis data afford a unique opportunity to examine this critical assumption empirically using causal studies that examine impacts on both instruction and achievement.

We take a straightforward approach to examining this hypothesis by estimating the correlation between coaching effects on instruction and effects on achievement from studies that estimated both (n=11).[7] First, we averaged effect size estimates for each outcome within a study. Then, we conducted a weight-based analysis using the average inverse variance of estimates for achievement outcomes.[8] Although we can interpret the effect of coaching on instruction and achievement in a causal framework, we cannot do so for the relationship between instruction and achievement. Our theory of change posits that improvements in instruction cause student achievement to rise. However, it is also possible that coaching effects on achievement were mediated through avenues other than instructional improvement (e.g., preparation time out of class). As such, we view this analysis as exploratory in nature. Access to the original data from these studies would allow us to instrument for instructional measures via random assignment of coaching, and we encourage future studies to engage in this type of analysis.

[7] These studies are denoted with a ^ in the references.

[8] Weighting results using variance estimates from instructional effect sizes produces qualitatively similar results.

Across our analyses we find strong supporting evidence for the link between instruction and achievement. Across a small sample of 11 data points, the correlation between effect sizes on instruction and achievement is .57 (*p* = .07; see also Figure 3). In addition to asking how effect sizes on instruction and achievement covary, we can interpret the magnitude of this relationship by examining how large of a change in achievement is associated with a given change in instruction. Here, we find that changes in student achievement appear to require relatively large improvements in instructional quality. Using a simple linear regression framework, we estimate that a 1 SD change in instruction is associated with a .33 SD change in achievement (*p* = .07). The reason that this point estimate is different from the correlation above is because the combined set of effect-size estimates is not standardized (this is illustrated by Figure 2). We also show that a linear projection of this relationship may not hold at higher levels of instructional effects in Figure 3 by overlaying a fitted quadratic function. This finding is consistent with a large body of literature documenting the weak relationship between educational inputs (instruction) and outputs (achievement) and helps to explain why PD that results in more modest changes in teachers' instruction often does not lead to impacts on student achievement.

Sensitivity Analyses

We examine the sensitivity of our estimates to three threats to internal validity: study design and data quality, outliers, and missing data. Beginning first with study design, we see few threats to internal validity given our strict inclusion criteria of studies capable of supporting causal inferences. The vast majority of studies are randomized control trials that are considered the gold standard of causal inference design (Murnane & Willett, 2011). Additional studies that met our inclusion criteria but used quasi-experimental designs all focused on a difference-in-difference strategy that rests on two critical assumptions: parallel trends between treatment and comparison groups, and no simultaneous confounding of treatment effects (Murnane and Willett, 2011). Given limited information to assess these assumptions directly, we instead probe the sensitivity of our findings to design and data quality by restricting the sample to only include randomized control trials. Unsurprisingly, our results remain quite similar to our main results when we exclude the four studies using difference-in-differences designs, with pooled effects of coaching on instruction of .54 SD and achievement of .15 SD (see Appendix Table A4).

Given the large variation in effect sizes as depicted in Figure 2, it is possible that our results are driven by outliers. Visual inspection of the data as well as box and whisker plots suggest there exist few clear outliers in our data. Rather than make a subjective decision about what data points constitute outliers, we test the sensitivity of our result by removing the lowest and highest 5% of our estimated effect sizes for each outcome. As shown in Appendix Table A5, our results are not driven by extreme values and remain largely unchanged after trimming the bottom and top 5% of estimates. We find pooled effects across all studies of .56 SD for instruction and .14 for achievement.

We also examine the degree to which our results may be a product of publication bias or non-reported outcomes. Data may be missing from the analytic dataset when studies that do not find statistically significant effects are not submitted or not accepted for publication, as well as when authors of published studies do not include the results of all available outcomes in a paper. We test the sensitivity of these findings by conducting a modified version of Duval and Tweedie's (2000) trim and fill method to account for the clustered nature of the data and the diverse range of coaching models in the analytic sample. Using this rank-based data augmentation technique, we estimate the number of missing effect sizes and impute these theoretically missing data points. This involves calculating the hypothetical data points needed to balance the spread of effect sizes across a centering estimate derived from the random effects model in equation 2. We do this first at the effect-size level by imposing a nested structure on the imputed data based on the average number of effect sizes per study in the analytic sample. We also replicate this approach after collapsing the data to the study level by averaging effect sizes and variance estimates within studies for a given outcome. As reported in Table 4, the adjusted estimates are attenuated, particularly for instructional outcomes, but remain statistically significant across both approaches. Pooled effect-size estimates are approximately .41 SD for instructional outcomes and .12 SD for achievement outcomes. These results suggest that our conclusions around the effectiveness of teacher coaching as a PD tool are unlikely to be driven by missing data.

Discussion

In order to interpret the substantive significance of our findings, we consider several benchmarks described by Hill, Bloom, Black, and Lipsey (2008) and Lipsey et al. (2012): the observed effect of similar interventions, policy-relevant performance gaps, normative expectations for students' academic growth, and cost. Our estimates of the effect of coaching on teachers' instructional practice (.58 SD) are larger than differences in measures of instructional quality between novice and veteran teachers' (.2 to .4 SD; Hill et al., 2015). Effects on students' academic performance (.15 SD) are of similar or larger magnitude than estimates of the degree to which teachers' improve their ability to raise student achievement during the first five to ten years of their careers, with estimates ranging from .05 to .15 SD (Atteberry, Loeb, & Wykoff 2015; Papay & Kraft, 2015). Effects on achievement also are larger than pooled estimates from causal studies of almost all other school-based interventions reviewed by Fryer (2017) including student incentives, teacher pre-service training, merit-based pay, general PD, data-driven instruction, and extended learning time. Interventions of comparable effect sizes on achievement include comprehensive school reform (.1 to .2 SD, depending on the school reform model; Borman, Hewes, Overman, & Brown, 2003), oversubscribed charter schools (.04 SD to .08 SD per year of attendance; Chabrier, Cohodes, & Oreopoulos, 2016), large reductions in class size (roughly .2 SD; Krueger, 1999), high-dosage tutoring (.15 to .25 SD; Kraft, 2015; Blachman et al., 2004), and changes in curriculum (.05 to .3 SD depending on the grade level and curriculum under investigation; Agodini et al., 2009;Koedel, Li, Springer, & Tan, forthcoming).

From a policy perspective, the effects of teacher coaching must be considered relative to program costs. Traditional on-site coaching programs are a resource-intensive intervention simply due to the high personnel costs of staffing a skilled coaching corps. One cost analysis of coaching across three schools found per-teacher costs ranged from $3,300 to upwards of $5,200 (Knight, 2012). Unfortunately, the existing literature lacks the necessary information about program costs to conduct a reliable cost-benefit or cost-effectiveness analysis. As researchers and practitioners continue to innovate, they should explore ways to minimize costs while maintaining the efficacy of coaching. We highlight some of these possibilities, including virtual coaching, in the remaining part of our discussion and conclusion. However, if an instructional expert working one-on-one with teachers in person over a sustained amount of time remains at the core of effective coaching models, then this approach will always require fairly sizeable financial and human capital investments. Given the billions of dollars districts currently spend on PD, coaching should not be seen as prohibitively expensive from a policy perspective. Instead, policymakers and administrators must judge whether their current expenditures on PD could be maximized more effectively. One approach would be to allocate resources to high-cost but effective PD programs for teachers most in need of support, such as coaching, rather than to lower-cost but less-effective programs for all teachers.

Taking Teacher Coaching to Scale

Decades worth of research have documented the significant challenges of taking education programs and reform initiatives to scale (Honig, 2006). Given the fundamental importance of implementation quality, major questions still remain about the feasibility of expanding teacher coaching across schools and districts. For example, a literacy PD program modified for scalability by reducing coaching intensity, using trained research assistants as coaches, and providing only written feedback found no effects on children's language skills (Cabell et al., 2011). We first explore this question graphically by illustrating the relationship between teacher sample size and effects sizes in Figure 4. This figure depicts a scatterplot of the average effect size by ventiles of teacher sample size with the linear relationship from an OLS regression overlaid on top. Graphs for both instruction (Panel A) and achievement (Panel B) depict a clear negative relationship between the size of a coaching program and program effects.

We more formally test for evidence of potential scale-up implementation challenges by dividing the sample of studies into two groups following Wayne et al. (2008): *efficacy* trials (studies with samples of fewer than 100 teachers) versus *effectiveness* trials (studies with samples of 100 teachers or more). We use teacher sample size to provide a simple proxy measure for categorizing the studies while recognizing that some larger studies may share features of efficacy trials and vice versa. Efficacy trials examine small programs under conditions that are intended to be as conducive as possible to maximizing effects. Experimental studies with fewer than 100 teachers generally involved coaching no more than 50 teachers and required only a handful of coaches to implement. Typically, these studies evaluated the potential of coaching models under best-case conditions with researchers often playing a role in designing and delivering the coaching program to a small group of motivated volunteer teachers (e.g., Allen, Hafen, Gregory, Mikami, & Pianta, 2015; Matsumara et al.,2012; McCollum, Hemmeter, & Hsieh, 2013). Such programs often are tailored specifically for participating teachers and the school contexts in which they work. In contrast, larger-scale effectiveness trials test programs implemented at scale across a range of settings with more limited support. In our sample, effectiveness trials generally required recruiting and training a larger coaching corps to deliver a more standardized program across a broader range of contexts where teachers were more likely to have mixed levels of interest in the program (e.g., Garet et al., 2008, 2011; Lockwood et al., 2010).

Comparing pooled effect sizes estimates for efficacy versus effectiveness trails suggests that coaching can have an impact at scale but that scale-up implementation challenges likely attenuate this effect. As reported in Table 5, we estimate that smaller coaching programs improved classroom instruction by .72 SD and raised student achievement by .21 SD. These pooled effect sizes are approximately 1.5 times the size of effects on instruction for smaller programs and 2 times the size for achievement (.47 SD for instruction and .11 SD for achievement). The difference in effect size estimates is marginally significant for achievement (*p*=.06) but not for instruction (*p*=.12). Publication bias may explain some of this difference if efficacy trials with smaller effect sizes are less likely to be published due to a lack of statistical significance. Many of the larger effectiveness trials are institute reports funded by IES that are published online whether or not findings are statistically significant. At the same time, this difference is qualitatively large enough to conclude that scaling-up coaching programs introduces additional challenges to those confronted by small-scale demonstration models.

One primary implementation challenge is building a corps of capable coaches whose expertise is well matched to the diverse needs of teachers in a school or district. Blazar and Kraft (2015) show that this is a challenge even in smaller efficacy trials. Leveraging turnover of coaches across two cohorts of an experimental evaluation, they found that coaches varied significantly in their effectiveness at improving teachers' instructional practice. A common approach to filling the demand for high-quality coaches is to tap expert local teachers. This strategy comes with the tradeoff of potentially removing highly-effective teachers from the classroom, but could be partially addressed with teachers taking on coaching responsibilities only part-time. A recent study found that pairing teachers with different strengths and weaknesses and encouraging them to coach each other is a promising strategy closely related to the coaching programs included in this analysis (Papay et al., 2016). Another approach taken by many districts has been to fold coaching into the observation component of new teacher evaluation systems. However, both theory (Herman & Baker, 2009) and case-study analyses (Kraft & Gilmour, 2016) suggest that having the same person serve as both coach and evaluator can undercut the trusting relationships needed between coaches and teachers and may result in superficial and infrequent feedback. Simply adding coaching responsibilities to administrators' existing responsibilities with little training or support is unlikely to result in intensive or sustained coaching.

Web-based virtual coaching offers one model for addressing the need for high-quality coaches amidst resource constraints. Leveraging video-based technology can increase the number of teachers with whom an individual coach can work and has the potential to increase access to high-quality coaches for schools or districts without local expertise. This approach may also reduce reservations among teachers about having their coach also be their evaluator, as the coach is both physically separate from and unaffiliated with their school. Further, virtual coaching could lower coaching costs by eliminating commute time. The lack of any differences in effect sizes between in-person and virtual coaching suggests that virtual coaching models may be able to maintain quality while increasing scalability. This finding is consistent with Powell et al. (2010) who did not find any consistent differences in outcomes across teachers randomly assigned to an in-person coach versus a coach who met with teachers virtually.

The need for teacher buy-in presents a second major challenge for scaling-up coaching programs. No matter the expertise or enthusiasm of a coach, coaching is unlikely to impact instructional practice if the teachers themselves are not invested in or are uncomfortable with the coaching process. The programs included in this review likely benefit from the non-random sample of teachers and schools that volunteered to participate in the studies. The largest study in our sample points to the challenges of taking coaching to scale and potentially making participation mandatory. Lockwood et al. (2010) evaluate a statewide reading coaching program in Florida that ultimately employed over 2,300 coaches. Across the four years they studied, effects on student achievement in math were statistically significant in only one of the four years, and effects on reading achievement were statistically significant in only two of the four years. Across all years, average effect sizes were extremely small, between .01 SD and .03 SD. It is not possible to determine whether these results are due to the mandatory nature of the program or from the sheer size of these efforts and thus, the need for a large corps of coaches. However, this study points to the challenges of building effective coaching programs at scale for all teachers, including some of whom may not be open to participating in coaching.

The literature on schools as organizations provides some insights about how best to address the likely challenges of gaining teacher buy-in. Coaching requires teachers to be willing to open themselves to critique and recognize personal weaknesses. This openness on the part of teachers is facilitated both by a school culture committed to continuous improvement and by strong relational trust among administrators and staff members (Bryk & Schneider, 2002; Kraft & Papay, 2014). Teachers that perceive the observation and feedback cycles associated with teacher coaching as a process intended to document shortcomings towards efforts to exit teachers may be unwilling to acknowledge a coach's critiques or experiment with new techniques for fear that it may be used against them (Herman & Baker, 2009; Kraft & Gilmour, 2016). This suggests that building environments where providing and receiving constructive feedback is a regular part of teachers' professional work may be a key condition for the success of scale-up efforts.

Taking coaching programs to scale will require building an effective coaching corps as well as working with teachers with mixed levels of interest across schools with varying degrees of supportive school climates. There is no guarantee these challenges can be fully resolved. It may be that coaching is best utilized as a targeted program with a small corps of expert coaches working with willing participants rather than district-wide professional development.

Directions for Future Research

This systematic review of the literature also serves to identify important directions for future research. Most basically, we still know very little about the scope of teacher coaching programs as they currently are being implemented across the United States. We strongly encourage researchers to advocate for the inclusion of questions about coaching activities on nationally representative datasets such as the Schools and Staffing Survey and American Teacher Panel. The results also point to the relative lack of causal evidence on content-based coaching programs for subjects other than reading and literacy. The effect of coaching may differ across subject areas or for teachers with different levels of experience. Ongoing innovation in coaching practices is likely to produce new models which will present fertile areas for future research. One such example is “bug-in ear” coaching where peers stand in the back of a room and provide guidance to co-teachers in real-time via an earpiece (Scheeler, Congdon & Stansbery, 2010; Ottley, Coogle, Rahn, & Spear, 2017).

It also will be important to examine more closely which specific instructional practices are affected by coaching and which student outcomes improve as a result of these changes. Studies included in this analysis that measured instructional practice as an outcome tended to focus either on teachers’ literacy skills or teacher-student interactions as measured by instruments such as the CLASS. Sample size constraints for each type of teaching skill meant that we had to collapse all measures of teachers’ instructional practice into a single category. However, coaching may have differential impacts on different areas of teachers’ classroom practice, potentially driven by the theory of action of the coaching program itself or the skills of the coaches. In turn, different teaching skills have differential impacts on a range of student outcomes (e.g., academic achievement, behavior, self-efficacy; Blazar & Kraft, 2017). Understanding whether and how coaching can develop a broad range of teaching skills will be crucial in order to address the varied needs of teachers and students in classrooms across the U.S.

Similarly, we see a need for studies to move beyond efficacy trials to evaluate specific program design features, particularly those features that may be necessary to take programs to scale. Studies that randomize teachers or schools to coaching programs that differ by, for example, the number of coaching sessions, or in-person versus virtual coaching would be particularly informative. In cases where efficacy trials have demonstrated the potential of coaching models, such as with literacy coaching, researchers should turn towards evaluating these models in large-scale effectiveness trials where the evaluators are not primarily responsible for program implementation. Identifying the features of effective coaching programs and building the knowledge base about how to scale up such programs are, in our view, the most important areas for future research.

Finally, all futures studies would benefit from examining outcomes in the year after the coaching program ends. Among the 44 studies we reviewed, only four reported outcomes from a follow-up year after coaching had ended (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Blazar & Kraft, 2015; Garet et al., 2008; Teemant, 2014). These studies present very mixed evidence about the degree to which effects are enhanced, sustained, or fade out over time. Understanding the degree to which teachers continue to implement the practices they learned with the support of a coach is essential to considering the overall costs of rolling out coaching programs at scale. Admittedly, this is not always easy to do. Maintaining the internal validity of an experimental study over time can be challenging given high rates of teacher turnover, especially in urban large districts. Analytic methods, such as computing bounds on estimates (e.g., Lee, 2009) and tracking reasons for exiting a study, can help to address this challenge.

Inconsistencies in the reporting, design, and analysis of the existing literature of teaching coaching point to a need for researchers to strengthen the quality of future studies. Our ability to analyze specific features of coaching programs was limited by the lack of basic information available in many studies. This was particularly true for teacher and coach characteristics which are important for understanding who benefits from coaching and the background and training of effective coaches. Among the studies we reviewed that provided information about coaches, we found that coaches' had varied backgrounds including retired or master teachers affiliated with participating schools, university professors or graduate students with relevant teaching experience, and full-time coaches external to the district brought in by researchers.

We recommend researchers make it standard practice to collect and report the following information in as much detail as possible:

- The theory of action underpinning the coaching program
- The target population of teachers, including novice versus more veteran teachers
- The fidelity of implementation of the coaching model
- The length, frequency, and total amount of coaching sessions
- The length and features of other complementary PD elements of a coaching model
- Information on how teachers and schools were recruited and compare to those that did not volunteer for a study
- The number of coaches as well as any training and support they receive
- Coach background characteristics (e.g., teaching and coaching experience, subject expertise, role in school or district).
- Estimates of the per-teacher cost of delivering the coaching program
- A clear explanation of the type of PD available to teachers and schools in the control condition
- Information about the reliability of outcome measures including observation instruments, achievement tests and self-report surveys

This information will help to inform the research design process as well as provide essential information to researchers and practitioners interested in replicating or adopting these models.

Given the rigorous methodological inclusion criteria, the studies included in this review were overwhelmingly of high overall quality. However, there were several design and analysis practices that researchers could improve on in future studies. Many of the studies we reviewed were substantially underpowered to detect plausible effect sizes on distal outcomes such as student achievement. Studies would often have benefitted from randomizing at the teacher level instead of the school or district level. While this approach has disadvantages such as increasing the likelihood of spillover effects and limiting the opportunities for peer learning and support, we see the benefits of increased power as far outweighing these drawbacks (see Rhoads, 2011). Studies also could have been more consistent in collecting baseline measures of outcomes and other covariates that can serve to increase the precision of estimates. We also found examples of studies that did not properly account for the clustered nature of the data or the level of randomization when modeling standard errors. Finally, rates of attrition differed across studies in meaningful ways, while not all researchers tested for differential attrition or subjected their results to robustness checks for this attrition. Future reviews may consider coding studies based on these elements of research quality as well.

Conclusion

By pooling results from across 44 causal studies of teacher coaching, we find large positive effects on instruction and smaller positive effects on achievement. Effects on instruction and achievement compare favorably when contrasted with the larger body of literature on teacher PD (Yoon et al., 2007), as well as most other school-based interventions (Fryer, 2016). The growing literature on teacher coaching provides a much needed evidentiary base for future directions in teacher development policy, practice, and research. Ultimately, improving the teacher workforce in the U.S. will require continued innovation in in-service professional development programs such as teacher coaching given its vast size. Teacher coaching models can provide a flexible blueprint for these efforts but many questions remain about how best to implement these models at scale in a cost-effective manner.

References

* Indicates if a reference was included in the meta-analytic sample
^ Studies with both instruction and achievement outcomes that are included in Figure 3

*Abry, T., Rimm-Kaufman, S. E., Larsen, R. A., & Brewer, A. J. (2013). The influence of fidelity of implementation on teacher–student interaction quality in the context of a randomized controlled trial of the Responsive Classroom approach. *Journal of School Psychology*, *51*(4), 437-453. doi: 10.1016/j.jsp.2013.03.001.

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., & Murphy, R. (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools*. *NCEE 2009-4052.* National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/.

*Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(4), 475-489. doi: 10.1080/19345747.2015.1017680.

^*Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science,* *333*, 1034-1037. doi: 10.1126/science.1207998.

Amendum, S. J., Vernon-Feagans, L., & Ginsberg, M. (2011). The effectiveness of a technologically facilitated classroom-based early reading intervention. *The Elementary School Journal*, *112*, 107–131. doi: 10.1086/660684.

Angrist, J. D. (2004). American education research changes tack. *Oxford review of economic policy*, *20*(2), 198-212. doi: 10.1093/oxrep/grh011.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, *1*(4), 1–23. doi:10.1177/2332858415607834.

*Biancarosa, G., Bryk, A., & Dexter, E. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal*, *111*(1), 7-34. Retrieved from http://www.journals.uchicago.edu/

*Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A…. & Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI Program. Child Development, 79(6), 1802-1817. Retrieved from http://www.srcd.org/

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., & Shaywitz, S. E. (2004). Effects of Intensive Reading Remediation for Second and Third Graders and a 1-Year Follow-Up. *Journal of Educational Psychology*, *96*(3), 444. doi:10.1037/0022-0663.96.3.444.

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, *22*(2), 71-94.

*Blazar, D. & Kraft, M. A. (2015). Exploring Mechanisms of Effective Teacher Coaching: ATale of Two Cohorts From a *Randomized Experiment. Educational Evaluation and Policy Analysis, 37 (4),* 542–566. doi: 10.3102/0162373715579487

Blazar, D. & Kraft, M. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39(1),* 146-170.

*Boller, K., Del Grosso, P., Blair, R., Jolly, Y., Fortson, K., Paulsell, D.… & Kovas, M.. (2010). The seeds to success modified field test: Findings from the impact and implementation studies. *Mathematica Policy Research*. Retrieved from https://www.mathematica-mpr.com.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125-230. doi: 10.3102/00346543073002125.

Bowne, J. B., Yoshikawa, H., & Snow, C. E. (2016). Experimental impacts of a teacher professional development program in early childhood on explicit vocabulary instruction across the curriculum. *Early Childhood Research Quarterly*, *34*, 27-39. doi: 10.1016/j.ecresq.2015.08.002.

Bryk, A. & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York, NY: Russell Sage Foundation.

Cabell, S. Q., Justice, L. M., Piasta, S. B., Curenton, S. M., Wiggins, A., Turnbull, K. P., & Petscher, Y. (2011). The impact of teacher responsivity education on preschoolers' language and literacy skills. *American Journal of Speech-Language Pathology*, *20*(4), 315-330. doi:10.1044/1058-0360.

*Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*(3), 430-454. doi: 10.1086/657654.

Chabrier, J., Cohodes, S., & Oreopoulos, P. (2016). What can we learn from charter school lotteries?. *The Journal of Economic Perspectives*, *30*(3), 57-84. doi: 10.1257/jep.30.3.57.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*(9), 2633-2679. doi: 10.1257/aer.104.9.2633.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance:

The mathematics reform in California. Teachers College Record, *102*(2), 294–343.doi: 10.1111/0161-4681.00057

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational evaluation and policy analysis*, *25*(2), 119-142. doi: 10.3102/01623737025002119

*Conroy, M. A., Sutherland, K. S., Algina, J. J., Wilson, R. E., Martinez, J. R., & Whalon, K. J. (2015). Measuring teacher implementation of the BEST in CLASS intervention program and corollary child outcomes. *Journal of Emotional and Behavioral Disorders*, *23*(3) 1-12. doi:10.1177/1063426614532949.

Cook, T. D. (2001). Sciencephobia. *Education Next*, *1*(3). Retrieved from http://educationnext.org/.

Cornett, J., & Knight, J. (2009). Research on coaching. In *Coaching: Approaches and perspectives (*pp. 192-216). Thousand Oaks, CA: Corwin Press.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad.* Palo Alto, CA: National Staff Development Council and The School Redesign Network, Stanford University.

Denton, C. A., & Hasbrouck, J. A. N. (2009). A description of instructional coaching and its relationship to consultation. *Journal of Educational & Psychological Consultation*, *19 (2),* 150-175. doi: 10.1080/10474410802463296

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational researcher*, *38*(3), 181-199. doi: 10.3102/0013189X08331140.

Desimone, L. M., & Garet, M. S. (2015). Best practices in teachers' professional development in the United States. *Psychology, Society and Education*, *7*(3), 252-263.

Devine, M., Meyers, R., & Houssemand, C. (2013). How can coaching make a positive impact within educational settings?. *Procedia-Social and Behavioral Sciences*, *93*, 1382-1389. doi: 10.1016/j.sbspro.2013.10.048

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A. M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, *87*(2), 243-282. doi: 10.3102/0034654316687036.

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

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta—analysis. *Biometrics*, *56* (2), 455-463. doi: 10.1111/j.0006-341X.2000.00455.x.

*Fisher, D., Frey, N., & Lapp, D. (2011). Coaching middle-level teachers to think aloud improves comprehension instruction and student reading achievement. *The Teacher Educator*, *46*(3), 231-243. doi: 10.1080/08878730.2011.580043

Fletcher, S., & Mullen, C. A. (Eds.). (2012). *Sage handbook of mentoring and coaching in education*. Thousands Oaks, CA: Sage.

Fryer Jr, R. G. (2017). The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. In E. Duflo & A. Banerjee (Eds.) Handbook of Field Experiments. Vol. 2. (pp. 95-322). Amsterdam: North-Holland.

Gallucci, C., Van Lare, M. D., Yoon, I. H., & Boatright, B. (2010). Instructional coaching building theory about the role and organizational support for professional learning. *American Educational Research Journal*, *47*(4), 919-963. doi: 10.3102/0002831210371497

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. doi: 10.3102/00028312038004915

^*Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W…. & Sztejnberg, L. (2008). The impact of two professional development interventions on early reading instruction and achievement (NCEE 2008-4030). Washington, D.C.: *National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.*

^*Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K.… & Doolittle, F. (2011). Middle school mathematics professional development impact study: Findings after the second year of implementation (NCEE 2011-4024). Washington, DC: *National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.*

Garet, M. S., Heppen, J.B., Walters, K., Parkinson, J., Smith, T.M., ..., Wei, T.E. (2016,). Focusing on Mathematical Knowledge: The Impact of Content-Intensive Teacher Professional Development. (NCEE 2016-4010). Washington DC: *National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.*

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study*. (NCEE 2010-4027). Washington, DC: *National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.*

*Gregory, A., Allen, J., Mikami, A., Hafen, C., & Pianta, R. (2014). Effects of a professional development program on behavioral engagement of students in middle and high school. *Psychology in the Schools, 51*(2). doi: 10.1002/pits.21741

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., ... & Brackett, M. A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*(4), 461-487. doi: 10.1086/669616.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466-479. Retrieved from https://www.journals.elsevier.com.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics*, *95*(7), 798-812. doi: 10.1016/j.jpubeco.2010.11.009.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods*, *1*(1), 39-65. doi: 10.1002/jrsm.5

*Hemmeter, M. L., Snyder, P. A., Fox, L., & Algina, J. (2016). Evaluating the implementation of the Pyramid Model for promoting social-emotional competence in early childhood classrooms. *Topics in Early Childhood Special Education*, *36*(3), 133-146. doi: 10.1177/0271121416653386

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172-177. doi: 10.1111/j.1750-8606.2008.00061.x

Hill, H. C. (2007). Learning in the teacher workforce. *Future of Children, 17*(1), 111-127. doi: 10.1353/foc.2007.0004.

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research consensus, crossroads, and challenges. *Educational Researcher*, *42*(9), 476-487. doi: 10.3102/0013189X13512674.

Hill, H. C., Blazar, D., & Lynch, K. (2015). Resources for teaching: Examining personal and institutional predictors of high-quality instruction. *AERA Open*, *1*(4), 1-23.

Herman J. L., & Baker, E. L. (2009). Assessment policy: Making sense of the Babel. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of educational policy research* (pp. 176-190). New York: Routledge.

Honig, M. I. (2006). *New directions in education policy implementation*. Albany, NY: SUNY Press.

Ippolito, J. (2010). Three ways that literacy coaches balance responsive and directive relationships with teachers. *The Elementary School Journal*, *111*(1), 164-190. doi: 10.1086/653474

Jackson, C. K. (2016). *What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes* (Working Paper No. w22226). Cambridge, MA: National Bureau of Economic Research.

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human*

Resources, *39*(1), 50-79. doi: 10.2307/3559005.

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

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, *85*(4), 512-552. doi: 10.3102/0034654314561338.

Joyce, B. R., & Showers, B. (1981). Transfer of training: the contribution of "coaching". *Journal of Education*, 163-172.

Joyce, B., & Showers, B. (1982). The coaching of teaching. *Educational leadership*, *40*(1), 4-10.

Joyce, B. R., & Showers, B. (2002). *Student achievement through staff development* (3rd edition). Alexandria, VA:ASCD.

Kennedy, M. M. (2016). How does professional development improve teaching?. *Review of Educational Research*, *86*(4), 945-980. doi: 10.3102/0034654315626800.

Knight, D. S. (2012). Assessing the cost of instructional coaching. *Journal of Education Finance*, *38*(1), 52-80 doi: 10.1353/jef.2012.0010

Koedel, C., Li, J., Springer, M.G., & Tan, L. (forthcoming). The Impact of Performance Ratings on Job Satisfaction for Public School Teachers. *American Educational Research Journal*.

*Kraft, M.A. & Blazar, D. (forthcoming). Individualized Coaching to Improve Teacher Practice Across Grades and Subjects: New Experimental Evidence. *Education Policy*. doi:10.1177/0895904816631099

Kraft, M.A. & Papay, J.P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*. *36*(4), 476-500

Kraft, M.A. & Gilmour, A. (2016) Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711-753.

Kretlow, A. G., & Bartholomew, C. C. (2010). Using coaching to improve the fidelity of evidence-based practices: A review of studies. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497-532. doi: 10.1162/003355399556052.

*Landry, S. H., Anthony, J. L., Swank, P. R., & Monseque-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology*, *101*(2), 448. doi: 10.1037/a0013842.

^*Landry, S. H., Swank, P. R., Anthony, J. L., & Assel, M. A. (2011). An experimental study evaluating professional development activities within a state funded pre-kindergarten program. *Reading and Writing*, *24*(8), 971-1010. doi: 10.1007/s11145-010-9243-1.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, *76*(3), 1071-1102. doi: 10.1002/jae.2473.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*. (NCSER 2013-3000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

*Lockwood, J. R., McCombs, J. S., & Marsh, J. (2010). Linking reading coaches and student achievement: Evidence from Florida middle schools. *Educational Evaluation and Policy Analysis, 32*(3), 372-388. doi: 10.3102/0162373710373388

Lofthouse, R., Leat, D., Towler, C., Hallet, E., & Cummings, C. (2010). Improving coaching: evolution not revolution, research report. Education Trust. Retrieved from http://www.ncl.ac.uk/

Marsh, J. A., McCombs, J. S., Lockwood, J. R., Martorell, F., Gershwin, D., Naftel, S., . . .Crego, A. (2008). Supporting literacy across the Sunshine State: A study of Florida middle school reading coaches. Santa Monica, CA: RAND.

*Mashburn, A. J., Downer, J. T., & Hamre, B. K. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science, 14*(4), 179-196. doi: 10.1080/10888691.2010.516187

*Matsumara, L. C., Garnier, H. E., Correnti, R., Junker, B., & Bickel, D. D. (2010). Investigating the effectiveness of a comprehensive literacy coaching program in schools with high teacher mobility. *The Elementary School Journal, 111*(1), 35-62. doi: 10.1086/653469.

*Matsumara, L. C., Garnier, H. E., & Spybrook, J. (2012). The effect of content-focused coaching on the quality of classroom text discussions. *Journal of Teacher Education, 63*(3), 214-228. doi: 10.1177/0022487111434985

^*Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction, 25,* 35-48. doi: 10.1016/j.learninstruc.2012.11.001.

*McCollum, J., Hemmeter, M., & Hsieh, W. (2013). Coaching teachers for emergent literacy instruction using performance based feedback. *Topics in Early Childhood Special*

Education, 33(1), 28-37. doi: 10.1177/0271121411431003

*Mikami, A. Y., Gregory, A., Allen, J. P., Pianta, R. C., & Lun, J. (2011). Effects of a teacher professional development intervention on peer relationships in secondary classrooms. *School Psychology Review, 40*, 367-385. Retrieved from http://naspjournals.org/loi/spsr

*Milburn, T. F., Girolametto, L., Weitzman, E., & Greenberg, J. (2014). Enhancing preschool educator's ability to facilitate conversations during shared book reading. *Journal of Early Childhood Literacy*, *14*(1), 105-140. doi: 10.1177/1468798413478261

Miles, K. H., Odden, A., Fermanich, M., Archibald, S., & Gallagher, A. (2004). Inside the black box of professional development spending: Lessons from comparing five urban districts. *Journal of Education Finance*, *30*(1), 1-26. doi: 10.3102/0013189X15580944.

*Morris, P., Mattera, S., Castells, N., Bangser, M., Bierman, K., & Raver, C. (2014). Impact findings from the Head Start CARES demonstration: National evaluation of three approaches to improving preschoolers' social and emotional competence. Washington, D.C.: *Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services*. Retrieved from http://www.acf.hhs.gov/opre

Murnane, R. J., & Nelson, R. R. (2007). Improving the performance of the education sector: The valuable, challenging, and limited role of random assignment evaluations. *Economics of Innovation and New Technology*, *16*(5), 307-322. doi: 10.1080/10438590600982236.

Murnane, R., & Willett, J. (2011). *Methods matter. Improving causal inference in educational and social science research*. Oxford, UK: Oxford University Press.

*Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Education*

Research Journal, 46(2), 532-566. doi: 10.3102/0002831208328088.

*Neuman, S. B., & Wright, T. S. (2010). Promoting language and literacy development for early childhood educators: A mixed-methods study of coursework and coaching. *The Elementary School Journal, 111*(1), 63-86. doi: 10.1.1.616.9207.

^*Nugent, G., Kunz, G., Houston, J., Kalutskaya, I., Wu, C., Pedersen, J…. & Berry, B. (2016). The effectiveness of technology-delivered science instructional coaching in middle and high school. *National Center for Research on Rural Education, Institute of Educational Sciences, U.S. Department of Education*.

Obara, S. (2010). Mathematics coaching: A new kind of professional development. *Teacher development, 14*(2), 241-251.doi: 10.1080/13664530.2010.494504.

Odden, A., Archibald, S., Fermanich, M., & Gallagher, H. A. (2002). A cost framework for professional development. *Journal of Education Finance*, *28*(1), 51-74.

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, *81*(3), 376-407. doi: 10.3102/0034654311413609.

Ottley, J. R., Coogle, C. G., Rahn, N. L., & Spear, C. F. (2017). Impact of Bug-in-Ear Professional Development on Early Childhood Co-Teachers' Use of Communication Strategies. *Topics in Early Childhood Special Education*, *36*(4), 218-229. doi: 10.1177/0271121416631123.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data* (Working Paper No. W21986). Cambridge, MA: National Bureau of Economic Research.

Papay, J.P. & Kraft, M.A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal* *of Public Economics,130*, 105-119.

^*Parkinson, J., Salinger, T., Meakin, J., & Smith, D. (2015). Results from a three-year i3 impact evaluation of the Children's Literacy Initiative (CLI): Implementation and impact findings of an intensive professional development and coaching program. *American Institutes for Research*. Retrieved from http://www.cli.org/.

*Pianta, R. C., Burchinal, M., Jamil, F. M., Sabol, T., Grimm, K., Hamre, B. K…. & Howes, C. (2014). A cross-lag analysis of longitudinal associations between preschool teachers' instructional support identification skills and observed behavior. *Early Childhood Research Quarterly, 29*, 144-154. doi: 10.1016/j.ecresq.2013.11.006.

*Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 23*, 431-451. doi: 10.1016/j.ecresq.2008.02.001

^*Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on Head Start teachers and children. *Journal of Educational Psychology, 102*(2), 299-312. doi: 10.1037/a0017763

Ramey, S. L., Crowell, N. A., Ramey, C. T., Grace, C., Timraz, N., & Davis, L. E. (2011). The dosage of professional development for early childhood professionals: How the amount and density of professional development may influence its effectiveness. *Advances in Early Education and Day Care*, 15, 11–32. doi: 10.1108/S0270-4021(2011)0000015005.

Randel, B., Beesley, A. D., Apthorp, H., Clark, T. F., Wang, X., Cicchinelli, L. F., & Williams, J. M. (2011). Classroom Assessment for Student Learning: Impact on Elementary School Mathematics in the Central Region. Final Report. (NCEE 2011-4005). Washington, DC:

National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Metzger, M. W., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: a cluster-randomized controlled trial. *Journal of consulting and clinical psychology*, *77*(2), 302. doi: 10.1037/a0015302.

Rezzonico, S., Hipfner-Boucher, K., Milburn, T., Weitzman, E., Greenberg, J., Pelletier, J., & Girolametto, L. (2015). Improving Preschool Educators' Interactive Shared Book Reading: Effects of Coaching in Professional Development. *American Journal of Speech-Language Pathology*, *24*(4), 717-732. doi: 10.1044/2015_AJSLP-14-0188.

Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics, 36*(1), 76-104. doi: 10.3102/1076998610379133.

Richard, A. 2003. 'Making our own road': The emergence of school-based staff developers in America's public schools. New York, NY: Edna McConnell Clark Foundation.

*Rimm-Kaufman, S. E., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., Merritt, E. G.… DeCoster, J. (2014). Efficacy of the Responsive Classroom Approach: Results from a 3-year, longitudinal randomized controlled trial. *American Educational Research Journal, 51*(3), 567-603. doi: 10.3102/0002831214523821.

Russo, A. 2004. School-based coaching: A revolution in professional development – Or just the latest fad? Harvard Education Letter. Retrieved from http://hepg.org/

Sailors, M., Hoffman, J. V., David Pearson, P., McClung, N., Shin, J., Phiri, L. M., & Saka, T. (2014). Supporting change in literacy instruction in Malawi. *Reading Research*

Quarterly, 49(2), 209-231. doi: 10.1002/rrq.70.

^*Sailors, M., Price, L. R. (2010). Professional development that supports the teaching of cognitive reading strategy instruction. *The Elementary School Journal, 110*(3), 301-322. doi: 10.3102/0162373715579487.

Sailors, M., & Shanklin, N. L. (2010). Introduction: Growing evidence to support coaching in literacy and mathematics. The Elementary School Journal, *111*(1), 1-6. doi: 10.1086/653467.

*Sailors, M., & Price, L. (2015). Support for the Improvement of Practices through Intensive Coaching (SIPIC): A model of coaching for improving reading instruction and reading achievement. *Teaching and Teacher Education*, *45*, 115-127. doi: 10.1016/j.tate.2014.09.008.

*Sibley, A., & Sewell, K. (2011). Can multidimensional professional development improve language and literacy instruction for young children? *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field, 14*(4), 263-274. doi: 10.1080/15240754.2011.609948

Schachter, R. E. (2015). An Analytic Study of the Professional Development Research in Early Childhood Education. *Early Education and Development, 26*(8), 1057-1085. doi: 10.1080/10409289.2015.1009335.

Scheeler, M. C., Congdon, M., & Stansbery, S. (2010). Providing immediate feedback to co-teachers through bug-in-ear technology: An effective method of peer coaching in inclusion classrooms. *Teacher Education and Special Education, 33*(1), 83-96. doi: 10.1177/0888406409357013.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental* *designs for generalized causal inference*. Boston, MA: Houghton, Mifflin and Company.

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. doi: 10.1080/19345740802641527.

Showers, B. (1984). Peer Coaching: A Strategy for Facilitating Transfer of Training. A CEPM R&D Report.Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Showers, B. (1985). Teachers coaching teachers. *Educational leadership*, *42*(7), 43-48.

Stormont, M., Reinke, W. M., Newcomer, L., Marchese, D., & Lewis, C. (2015). Coaching Teachers' Use of Social Behavior Interventions to Improve Children's Outcomes A Review of the Literature. *Journal of Positive Behavior Interventions*, *17*(2), 69-82. doi: 10.1177/1098300714550657.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling Complex Meta-analytic Data Structures Using Robust Variance Estimates: a Tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*(1), 85-112. doi: 10.1007/s40865-016-0026-5.

*Teemant, A. (2014). A mixed-methods investigation of instructional coaching for teachers of diverse learners. *Urban Education, 49*(5), 574-604. doi: 10.1177/0042085913481362

*Vernon-Feagans, L., Kainz, K., Hedrick, A., Ginsberg, M., & Amendum, S. (2013). Live webcam coaching to help early elementary classroom teachers provide effective literacy instruction for struggling readers: The targeted reading intervention. *Journal of Educational Psychology, 105*(4), 1175-1187. doi: 10.1037/a0032143

*Vogt, F., & Rogalla, M. (2009). Developing adaptive teaching competency through coaching. *Teacher and Teacher Education, 25*, 1051-1060. Retrieved from http://www.journals.elsevier.com

^*Wasik, B. A., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology*, *98*(1), 63. doi: 10.1037/0022-0663.98.1.63.

^*Wasik, B. A., & Hindman, A. H. (2011). Improving vocabulary and pre-literacy skills of at-risk preschoolers through teacher professional development. *Journal of Educational Psychology, 103*(2), 455-469. doi: 10.1037/a0023067

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational researcher*, *37*(8), 469-479.doi: 10.3102/0013189X08327154.

Weiss, I. R., & Miller, B. (2006, October). Deepening teacher content knowledge for teaching: a review of the evidence. Paper presented at the Second MSP Evaluation Summit, Washington, D.C.

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M., Weiland, C., ... & Arbour, M. C. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental psychology*, *51*(3), 309. doi: 10.1037/a0038785.

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

*Zan, B., & Donegan-Ritter, M. (2013). Reflecting, coaching and mentoring to enhance teacher-child interactions in Head Start classrooms. *Early Childhood Education Journal, 42,* 93-104. doi: 10.1007/s10643-013-0592-7

Tables

Table 1. Study Characteristics

	Count	Proportion
Source		
Institute Report	6	0.14
Peer-reviewed Journal	38	0.86
Year of Publication		0.00
2006	1	0.02
2008	3	0.07
2009	4	0.09
2010	8	0.18
2011	9	0.20
2012	1	0.02
2013	3	0.07
2014	7	0.16
2015	5	0.11
2016	2	0.05
in press	1	0.02
Research Design		
Randomized Control Trials (RCTs)	40	0.91
Quasi-experiment	4	0.09
Level of Randomization for RCTs		
Teacher	17	0.43
School	21	0.53
District	2	0.05
Teacher Sample Size		
50 or less	11	0.25
51 to 100	13	0.30
101 to 150	6	0.14
151 to 300	11	0.25
300 plus	2	0.05
Not reported	1	0.02
Coaching Model Type		
Content-Specific	29	0.66
Math	2	0.05
Reading	25	0.57
Science	2	0.05
General Practices	15	0.34
School Levels Included		
Pre-K	20	0.45
Elementary	16	0.36
Middle	13	0.30
High	6	0.14
Mode of Delivery		
In Person	35	0.80
Virtual	9	0.20
Complementary Treatment Elements		

Any Complementary Treatment	40	0.91
Group Trainings	37	0.84
Instructional Content	18	0.41
Video Library	8	0.18
Coaching Dosage (# of hours of one-on-one coaching)		
10 or less	9	0.20
11 to 20	11	0.25
21 to 30	5	0.11
30 or more	7	0.16
Not reported	12	0.27
Total PD Dosage (# of hours)		
20 or less	8	0.18
21 to 40	10	0.23
41 to 60	9	0.20
60 or more	9	0.20
Not reported	8	0.18
n	44	

Notes: School levels included is not mutually exclusive as several studies include sample of teachers from across schooling levels.

Table 2. Pooled Effect Size Estimates

	Classroom Observations	Achievement (Pooled)	Reading Achievement	Math Achievement	Science Achievement
All Studies	0.584***	0.147***	0.167***	0.022	0.111
	(0.065)	(0.027)	(0.033)	(0.044)	(0.025)
k[n]	155[32]	82[23]	60[19]	19[4]	3[2]
Content-Specific (All)	0.568***	0.158***			
	(0.063)	(0.026)	↓	na	na
k[n]	99[20]	77[20]			
Content-Specific (Reading)	0.577***	0.183***	0.183***		
	(0.067)	(0.032)	(0.032)	na	na
k[n]	93[18]	60[16]	56[16]		
General Practices	0.630**	0.096	0.101		
	(0.148)	(0.139)	(0.126)	na	na
k[n]	56[12]	5[3]	4[3]		

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies. The pooled estimate of the effect of content-specific (all) coaching programs on reading achievement is omitted because it is identical and uses the same sample as content-specific (reading) coaching programs on reading achievement. Cells with "na" are not estimated due to too few or no data.

Table 3. Pooled Effect Size Estimates by School Level

	Classroom Observations	Achievement (Pooled)
Pre-Kindergarten	0.662***	0.179***
	(0.075)	(0.040)
k[n]	120[18]	26[6]
Elementary School	0.581**	0.165***
	(0.181)	(0.043)
k[n]	21[9]	40[11]
Middle School	0.435***	0.087*
	(0.066)	(0.034)
k[n]	22[8]	19[9]
High School	0.471**	0.177*
	(0.156)	(0.080)
k[n]	15[3]	4[3]

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. Pre-Kindergarten coaching programs only have achievement outcomes for reading. For sample size, k is the number of effect sizes and n is the number of studies.

Table 4. Sensitivity Analyses using Modified Trim and Fill Method

	Effect-Size Level		Study Level	
	Classroom Observations	Achievement (Pooled)	Classroom Observations	Achievement (Pooled)
	Panel A: Unadjusted Estimates			
All studies	0.584***	0.147***	0.533***	0.144***
	(0.065)	(0.027)	(0.057)	(0.027)
k[n]	155[32]	82[23]	[32]	[23]
	Panel B: Estimates with Imputed Missing Studies			
All studies	0.409***	0.119***	0.411***	0.130***
	(0.091)	(0.029)	(0.066)	(0.027)
k[n]	191[39]	96[27]	[43]	[27]

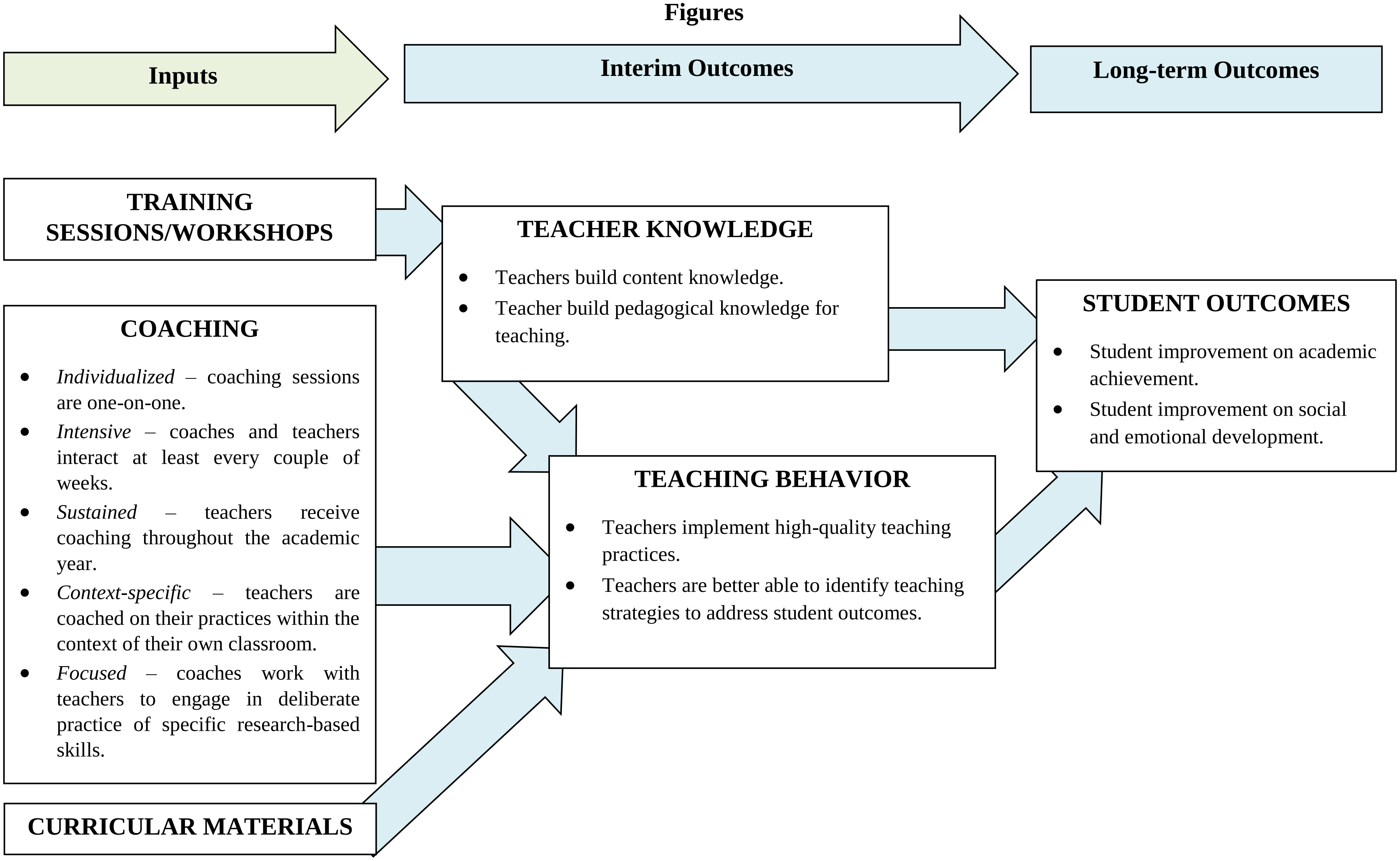
Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies. For effect-size level imputation we cluster effect sizes within studies according to the average number of effect-sizes per study in our primary samples.

Table 5. Pooled Effect Size Estimates by Coaching Program Size

	Classroom Observations	Achievement (Pooled)
All Studies	0.584***	0.147***
	(0.065)	(0.027)
k[n]	155[32]	82[23]
Efficacy Trials (n Teachers <100)	0.716***	0.212***
	(0.113)	(0.036)
k[n]	83	30
Effectiveness Trials (n Teachers ≥100)	0.472***	0.105**
	(0.072)	(0.037)
k[n]	72[14]	52[12]

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

Figures



Inputs

Interim Outcomes

Long-term Outcomes

TRAINING SESSIONS/WORKSHOPS

COACHING

- *Individualized* – coaching sessions are one-on-one.
- *Intensive* – coaches and teachers interact at least every couple of weeks.
- *Sustained* – teachers receive coaching throughout the academic year.
- *Context-specific* – teachers are coached on their practices within the context of their own classroom.
- *Focused* – coaches work with teachers to engage in deliberate practice of specific research-based skills.

CURRICULAR MATERIALS

TEACHER KNOWLEDGE

- Teachers build content knowledge.
- Teacher build pedagogical knowledge for teaching.

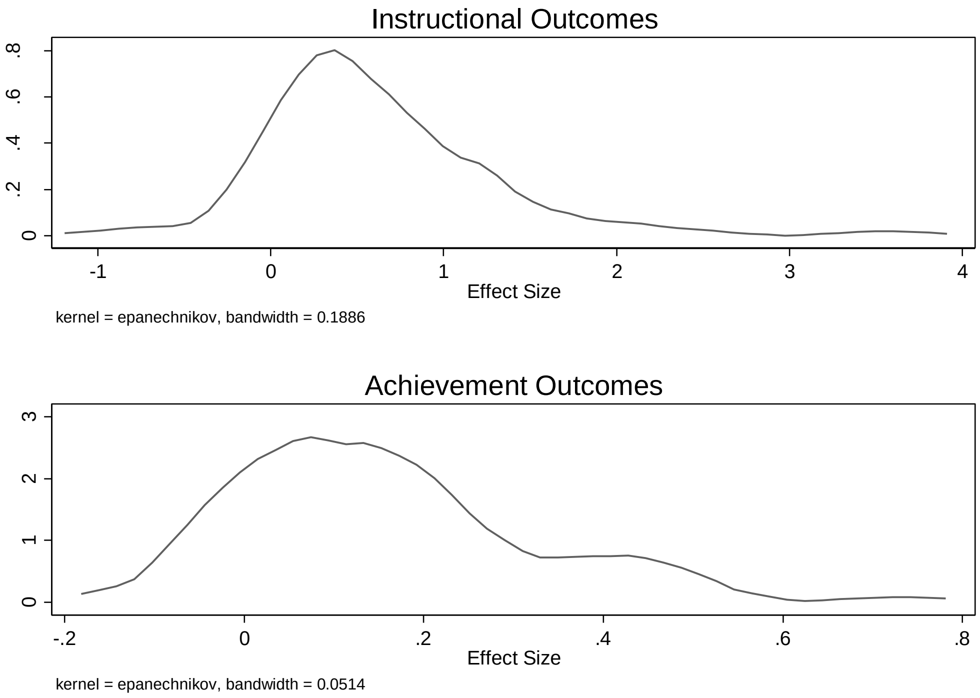
TEACHING BEHAVIOR

- Teachers implement high-quality teaching practices.
- Teachers are better able to identify teaching strategies to address student outcomes.

STUDENT OUTCOMES

- Student improvement on academic achievement.
- Student improvement on social and emotional development.

Figure 1: Theory of Action for Teaching Coaching


Instructional Outcomes

kernel = epanechnikov, bandwidth = 0.1886

Achievement Outcomes

kernel = epanechnikov, bandwidth = 0.0514

Figure 2: Kernel density plots of effect sizes for instructional and achievement outcomes.

Notes: k=155 for instructional outcomes and 82 for achievement outcomes

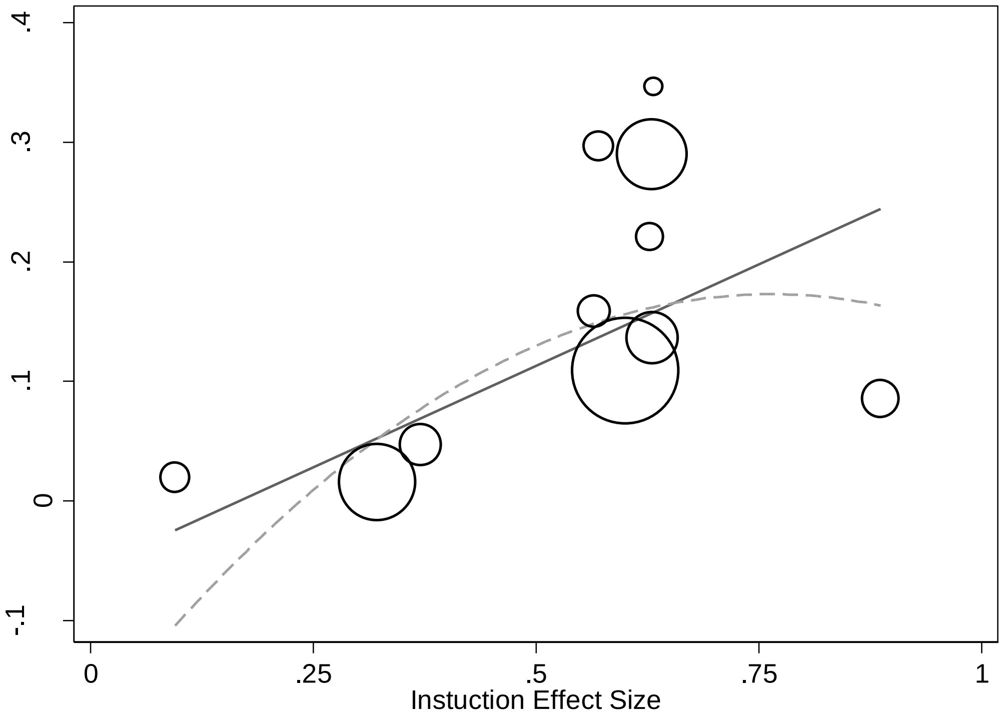
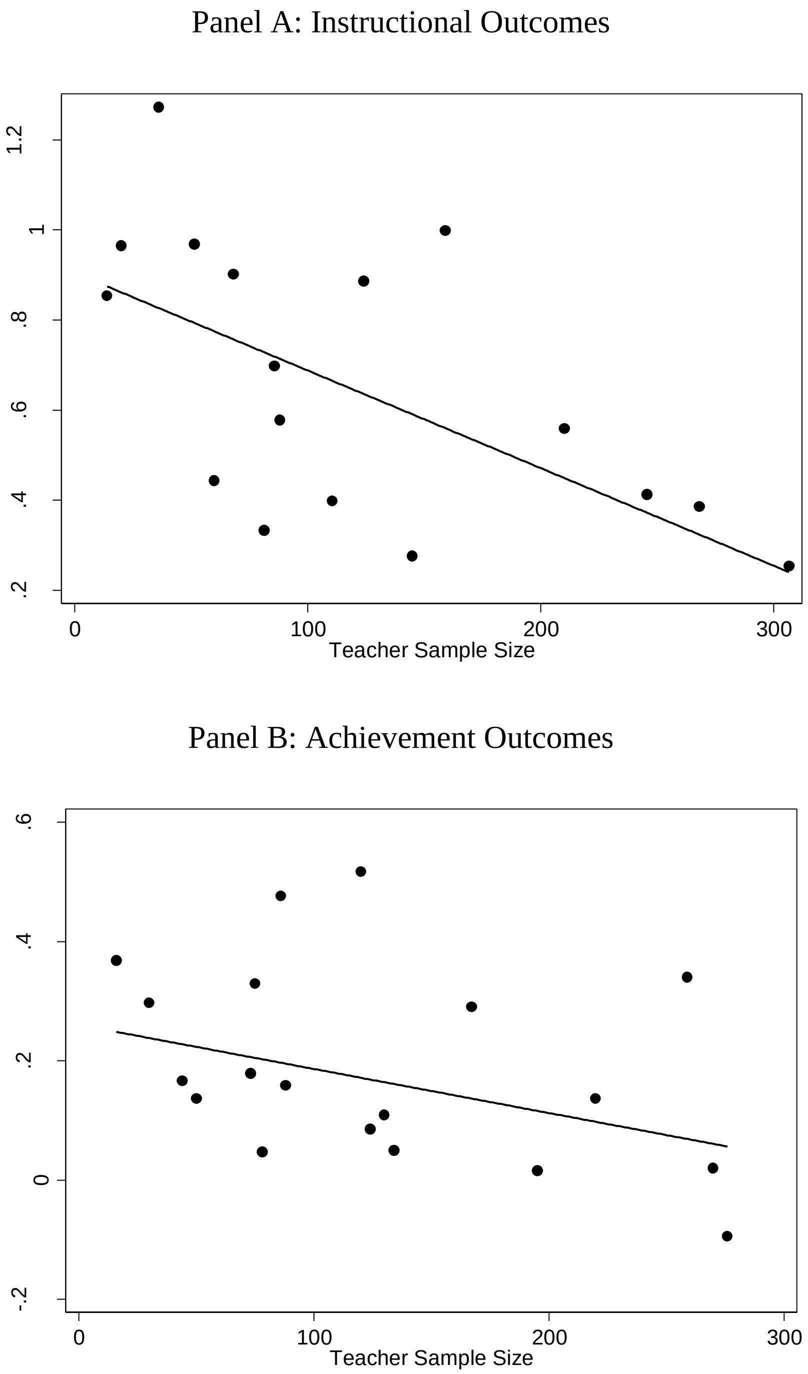


Figure 3. The relationship between coaching program effects on instruction and achievement.

Notes: Data points are calculated by averaging across effect sizes for a given outcome within studies and weighted by the product of the inverse of the average variance of achievement outcomes and instructional outcomes. n=11


Panel A: Instructional Outcomes

Panel B: Achievement Outcomes

Figure 4: The relationship between effect sizes and the number of teachers participating in a study

Notes: To construct these figures, we bin test scores into twenty equal sized (5 percentile point) bins and plot the mean effect size within each bin. The solid line shows the best linear fit estimated on the underlying data using OLS. Panel B excludes Cambell and Malkus et al. (2011) which reports a total teacher sample size of 1,593 and Lockwood et al. (2009) which does not report sample sizes for teachers.

Appendix Tables

TA 1. Studies Included in Meta-Analysis

Citation	Effective Teacher Sample Size	School Level	Research Design	Outcomes	Program Type	Complementary PD Features
Abry et al. (2013)	239	Elementary	RCT	Instruction	General Instruction	Group Training, Curriculum
Allen at al. (2015)	86	Middle, High	RCT	Achievement	General Instruction	Group Training, Video Library
Allen et al. (2011)	78	Middle	RCT	Instruction & Achievement	General Instruction	Group Training, Video Library
Biancarosa, Bryk, & Dexter (2010)	259	Elementary	Diff-in-diffs	Achievement	Reading Instruction	Group Training
Bierman et al. (2008)	44	Pre-K	RCT	Achievement	Reading Instruction & General Instruction	Group Training, Curriculum
Blazar & Kraft (2015)	82	Elementary, Middle, High	RCT	Instruction	General Instruction	Group Training, Curriculum
Boller et al. (2010)	159	Pre-K	RCT	Instruction	General Instruction	Group Training

Campell & Malkus (2011)	1593	Elementary	RCT	Achievement	Math Instruction	
Conroy et al. (2015)	53	Pre-K	RCT	Instruction	General Instruction	Group Training, Curriculum
Domitrovich et al. (2009)	84	Pre-K	RCT	Instruction	General Instruction	Group Training, Curriculum
Fisher, Frey, & Lapp (2011)	16	Middle	RCT	Achievement	Reading Instruction	Group Training
Garet et al. (2008)	270	Elementary	RCT	Instruction & Achievement	Reading Instruction	Group Training
Garet et al. (2011)	195	Middle	RCT	Instruction & Achievement	Math Instruction	Group Training
Gregory et al. (2014)	87	Middle, High	RCT	Instruction	General Instruction	Group Training, Video Library
Hemmeter et al. (2016)	40	Pre-K	RCT	Instruction	General Instruction	Group Training, Curriculum
Kraft & Blazar (in press)	50	Elementary, Middle, High	RCT	Instruction	General Instruction	Group Training, Curriculum

Landry et al. (2009)	262	Pre-K	RCT	Instruction	Reading Instruction	Group Training, Curriculum
Landry et al. (2011)	220	Pre-K	RCT	Instruction & Achievement	Reading Instruction	Group Training, Curriculum
Lockwood, McCombs, & Marsh (2010)		Middle	Diff-in-diffs	Achievement	Reading Instruction	
Mashburn et al. (2010)	134	Pre-K	RCT	Achievement	Reading Instruction & General Instruction	Curriculum, Video Library
Matsumara, Garnier, & Spybrook (2013)	167	Elementary	RCT	Instruction & Achievement	Reading Instruction	Group Training
Matsumara, Garnier, & Spybrook (2012)	93	Elementary	RCT	Instruction	Reading Instruction	Group Training
Matsumura et al. (2010)	73	Elementary	RCT	Achievement	Reading Instruction	
McCollum, Hemmeter, & Hsieh (2011)	13	Pre-K	RCT	Instruction	Reading Instruction	Group Training
Mikami et al. (2011)	88	Middle	RCT	Instruction	General Instruction	Group Training, Video Library

Milburn et al. (2014)	20	Pre-K	RCT	Instruction	Reading Instruction	Group Training, Curriculum
Morris et al. (2014)	308	Pre-K	RCT	Instruction	General Instruction	Group Training, Curriculum
Neuman & Cunningham (2009)	291	Pre-K	RCT	Instruction	Reading Instruction	Group Training
Neuman & Wright (2010)	148	Pre-K	RCT	Instruction	Reading Instruction	
Nugent et al. (2016)	124	Middle, High	RCT	Instruction & Achievement	Science instruction	Group Training, Curriculum
Parkinson et al. (2015)	130	Elementary	RCT	Instruction & Achievement	Reading Instruction	Group Training
Pianta et al. (2008)	113	Pre-K	RCT	Instruction	Reading Instruction & General Instruction	Curriculum, Video Library
Pianta et al. (2014)	252	Pre-K	RCT	Instruction	General Instruction	Video Library
Powell et al. (2010)	88	Pre-K	RCT	Instruction & Achievement	Reading Instruction	Group Training, Curriculum, Video Library

Rimm-Kaufman et al. (2014)	276	Elementary	RCT	Achievement	General Instruction	Group Training
Sailors & Price (2010)	44	Elementary, Middle	RCT	Instruction & Achievement	Reading Instruction	Group Training
Sailors & Price (2015)	120	Elementary, Middle	RCT	Achievement	Reading Instruction	Group Training
Sibley & Sewell (2011)	68	Pre-K	RCT	Instruction	Reading Instruction	Group Training, Curriculum
Teemant (2014)	36	Elementary	Diff-in-diffs	Instruction	General Instruction	Group Training
Vernon-Feagans et al. (2013)	75	Elementary	RCT	Achievement	Reading Instruction	Group Training
Vogt & Rogalla (2009)	50	Elementary, Middle, High	Diff-in-diffs	Achievement	Science instruction	Group Training
Wasik, Bond, & Hindman (2006)	16	Pre-K	RCT	Instruction & Achievement	Reading Instruction	Group Training, Curriculum
Wasik & Hindman (2011)	30	Pre-K	RCT	Instruction & Achievement	Reading Instruction	Group Training, Curriculum, Video Library

Zan & Donegan-Ritter (2014)	60	Pre-K	RCT	Instruction	Reading Instruction	Group Training

Table A2. Meta-regression Estimates of Coaching Program Moderators for Instruction Outcome

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Group Training	0.189			0.126						
	(0.142)			(0.195)						
Instructional Content		0.082		0.059						
		(0.133)		(0.142)						
Video Library			-0.161	-0.123						
			(0.111)	(0.139)						
Total # Complementary PD Features					0.030					
					(0.105)					
Virtual Coaching						-0.115				
						(0.122)				
Coaching Dosage							-0.002			
							(0.005)			
11-20 Coaching Hours								0.175		
								(0.136)		
21-30 Coaching Hours								-0.074		
								(0.171)		
31 or More Coaching Hours								0.042		
								(0.186)		
Total PD Dosage									-0.002	
									(0.003)	
21-40 Total PD Hours										0.253*
										(0.129)
41-60 Total PD Hours										0.077
										(0.295)
61 or More Total PD Hours										0.092
										(0.181)
Intercept	0.430***	0.555***	0.631***	0.494*	0.542**	0.619***	0.583***	0.512***	0.667***	0.488***
	(0.120)	(0.100)	(0.085)	(0.230)	(0.197)	(0.084)	(0.102)	(0.079)	(0.150)	(0.110)
k[n]	155[32]	155[32]	155[32]	155[32]	155[32]	155[32]	129[26]	129[26]	137[28]	137[28]

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

Table A3. Meta-regression Estimates of Coaching Program Moderators for Student Achievement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Group Training	0.076			0.074						
	(0.050)			(0.056)						
Instructional Content		0.022		0.017						
		(0.051)		(0.059)						
Video Library			-0.012	-0.009						
			(0.074)	(0.083)						
Total # Complementary PD Features					0.025					
					(0.026)					
Virtual Coaching						-0.000				
						(0.073)				
Coaching Dosage							-0.001			
							(0.002)			
11-20 Coaching Hours								-0.139+		
								(0.075)		
21-30 Coaching Hours								-0.117+		
								(0.063)		
31 or More Coaching Hours								-0.155+		
								(0.092)		
Total PD Dosage									-0.001	
									(0.001)	
21-40 Total PD Hours										0.110
										(0.073)
41-60 Total PD Hours										0.009
										(0.103)
61 or More Total PD Hours										-0.029
										(0.074)
Intercept	0.087*	0.140***	0.150***	0.086+	0.115*	0.148***	0.166***	0.239***	0.187***	0.123+
	(0.038)	(0.037)	(0.031)	(0.045)	(0.046)	(0.031)	(0.049)	(0.061)	(0.052)	(0.064)
k[n]	82[23]	82[23]	82[23]	82[23]	82[23]	82[23]	56[17]	56[17]	60[19]	60[19]

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

Table A4. Pooled Effect Size Estimates from Randomized Control Trials

	Classroom Observations	Achievement (Pooled)	Reading Achievement
All Studies	0.536***	0.146***	0.166***
	(0.051)	(0.029)	(0.033)
k[n]	153[31]	70[20]	53[17]
Content-Specific (All)	0.568***	0.158***	
	(0.063)	(0.027)	↓
k[n]	99[20]	65[17]	
Content-Specific (Reading)	0.577***		0.184***
	(0.067)	→	(0.030)
k[n]	93[18]		49[14]
General Practices	0.497***	0.096	0.101
	(0.087)	(0.139)	(0.126)
k[n]	54[11]	5[3]	4[3]

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies. Pooled estimates of the effect of content-specific (all) coaching programs on reading achievement and content-specific (reading) coaching programs on achievement (pooled) are omitted because they are identical and use the same sample as content-specific (reading) coaching programs on reading achievement.

Table A5. Pooled Effect Size Estimates after Trimming Top and Bottom 5% of Effect Sizes

	Classroom Observations	Achievement (Pooled)	Reading Achievement
All Studies	0.563***	0.139***	0.155***
	(0.054)	(0.026)	(0.032)
k[n]	139[32]	75[21]	54[17]
Content-Specific (All)	0.566***	0.156***	0.182***
	(0.059)	(0.027)	(0.032)
k[n]	91[20]	72[19]	51[15]
Content-Specific (Reading)	0.576***		0.182***
	(0.062)	→	(0.032)
k[n]	85[18]		51[15]
General Practices	0.569***		-0.005
	(0.107)	→	(0.054)
k[n]	48[12]		3[2]

Notes: * p<.05, ** p<.01, *** p<.001. Trimming top and bottom 5% of effect sizes removes 16 effect size estimates for the instruction sample and 8 effect size estimates for the achievement sample. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. The pooled estimate of content-specific (reading) coaching programs on achievement (pooled) is omitted because it is identical and uses the same sample as content-specific (reading) coaching programs on reading achievement. The pooled estimate of general practices coaching programs on achievement (pooled) is omitted because it is identical and uses the same sample as general practices coaching programs on reading achievement.