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

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 37 studies that employ causal research designs, we find pooled effect sizes of .57 standard deviations (SD) on instruction and .11 SD on achievement. Much of this evidence comes from studies of literacy coaching, which have an effect of .14 SD on reading achievement. Although these findings affirm the effectiveness 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 roughly 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 experimental 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:
http://scholar.harvard.edu/mkraft/publications/effect-teacher-coaching-instruction-and-achievement-meta-analysis-causal

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.

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 face unique challenges educating classrooms of students who enter with a wide range of prior 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 to rise and states adopt new content standards that require teachers to integrate higher-order thinking and social-emotional learning into the curriculum.

[1] Arriving at an exact estimate of total expenditures on PD is complicated by the fact that federal reporting 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).

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 et al., 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 et al., 2009; Desimone, 2009; Desimone & Garet, 2015; Garet et al., 2001; Hill, 2007). In this article, we review the evidence on a PD model that is centered on these core "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 in 2002, and the reauthorization of the Individuals with Disabilities Education Act in 2004. The legacy of these investments are 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 et al., 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. We focus our review narrowly on studies that employ research designs capable of supporting causal inferences.

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). This finding, along with substantial investments by the Institute for Education Research (IES) and a growing movement calling for increased methodological rigor 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 37 studies of teacher coaching programs that used both a causal research design and examined effects on instruction or student achievement (studies denoted with an * in the references). 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, by analyzing results across a range of coaching models, we are able to better understand the efficacy of coaching as a general class of PD. Second, the large financial and logistical costs of conducting experimental evaluations of teacher coaching programs has resulted in many underpowered individual studies. Meta-analysis techniques allow us to 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, pooling effects across studies allows us to compare different coaching models and examine specific design features that may drive program effects such as pairing coaching with other PD elements, in-person versus virtual coaching, and coaching dosage (Blazar & Kraft, 2015; Marsh et al., 2008; Ramey et al., 2011). Finally, subgroup analyses of larger versus smaller programs help identify implementation challenges and potential solutions for bringing coaching to scale.

Our analyses are driven by three 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?
>
> RQ3: What are some of the implementation challenges and potential opportunities for scaling up high-quality coaching programs in cost effective ways?

We pair empirical evidence from our meta-analytic strategy with a discussion of the broader literature and then conclude with recommendations on how future studies can strengthen and extend the causal research on teacher coaching. By examining these questions we hope to shed new 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.

Research Design

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, p. 1126; see also 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; 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 any teacher PD program is increased student learning (Devine et al., 2013; Desimone, 2009; Schachter, 2015), often operationalized somewhat narrowly as performance on standardized achievement tests. Mapping backwards, many argue that student achievement will not increase without changes in teacher knowledge or classroom practice (Cohen & Hill, 2000; Hanssen, 2006; Scher & O'Reilly, 2009). Training sessions, which are a standard form of PD offered to teachers (Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009; Hill, 2007), are thought to be beneficial in improving teachers' knowledge. However, additional levers are necessary for "facilitating the voluntary aspect of impacting teachers' disposition" (Schachter, 2015, p. 1067) and how they operationalize this knowledge through pedagogical practice in the classroom. Teacher coaching is considered a key lever to improve teachers' classroom instruction and to translate 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 may also 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.

For the purpose of this paper, we define coaching programs broadly as all PD programs that incorporate coaching as a key feature of the model. The role of the coach may be performed by a range of personell 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 our analyses that range from those focused on supporting the implementation of curriculum or pedagogical models to those where the coaching process itself is the core development tool.

Literature Search Procedures

We conducted a systematic review of the research literature through a three-phase process. We first identified articles by searching on electronic databases including Academic Search Premier, ERIC, PsycINFO, Ed Abstracts, ProQuest Theses and Dissertations, and Google Scholar. We used a variety of search terms such as "Teach* AND coach*" and then refined searches by combining these with additional terms such as "in-service", "professional development", "evaluation", "program effectiveness", "model*", "best practices". We then reviewed the references from existing literature reviews of coaching programs and references from the studies we identified in order to cross-check our search process. Finally, we contacted dozens of leading scholars in the field to solicit their help in identifying causal analyses of teacher coaching programs that we may have missed.

Inclusion Criteria

We restricted our sample of studies using four primary criteria pertaining to the sample, the intervention, the research design, and the outcomes. First, we limited our review to only include studies where the sample was comprised of early childhood to 12th grade teachers.

Second, we required that studies evaluate a PD program that included teacher coaching as a central feature following our working definition of coaching described above. Although we conceptualize coaching models as being both intensive and sustained, we chose not to set any minimum requirements on the frequency or duration of coaching. This less restrictive approach was motivated by our interest in examining empirically whether coaching effects differed by the intensity of the coaching model. We did, however, require that authors referred to the expert as a "coach" or the intervention as incorporating "coaching."

As our third criteria, we required that studies employed a research design capable of supporting causal inferences. This included studies that employed randomized control trials, as well as quasi-experimental methods such as regression discontinuity (Gamse et al., 2008), difference in differences (Teemant, 2014; Vogt & Rogalla, 2009), and multi-cohort longitudinal designs that modeled changes in achievement growth over time (Biancarosa, Bryk, & Dexter, 2010; Lockwood, McCombs, & Marsh, 2010). We excluded pre-post design studies that did not include a comparison group or relied principally on covariate adjustment given concerns that these strategies cannot adequately account for non-random selection (Shadish, Cook, & Campbell, 2002; Murnane & Willett, 2011).

Finally, 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.[2] When multiple

[2] 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.

papers were published using the same or overlapping set of experimental data we chose to include only one of the studies in our analysis (e.g., Blazar & Kraft, 2015 instead of Kraft & Blazar, in press; Vernon-Feagans et al., 2013 instead of Amendum, Vernon-Feagans, & Ginsberg, 2011).

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 our measures of instruction to only include those that were collected by outside observers blind to treatment status.[3] 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, 2015; Hamre et al., 2013), it was not possible to examine these constructs separately in our analyses. Studies used many different observation instruments and provided varying levels of information about these instruments, limiting our ability to assess the degree of overlap among specific dimensions. These instruments included observation rubrics that are well-established in the research literature and widely used by districts (e.g., Classroom Assessment Scoring System

[3] The number of observations per teacher varies considerably across studies. We do not impose a minimum number of observersations 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).

Student Achievement. We included in our analyses impacts on students' performance from a range of standardized achievement tests. These included both formative and summative assessments administered as part of the normal schooling process as well as those administered specifically for research purposes. Formative 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). Summative 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 our results by subject given the clear content distinctions across these measures.

Coding Procedures

We coded studies for effect sizes and associated standard errors as well as a range of study characteristics and coaching model features. In the few instances where sufficient information was not provided to calculate effect sizes (e.g., the standard deviation of the outcome measure), we were able to obtain the necessary information directly from the authors. We developed codes for 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 one author and then reviewed by a second author to ensure high reliability. Authors conferred and arrived at a consensus agreement when discrepancies arose. We describe the codes we used to characterize study features below:

Source and Year of Publication: We categorized the source of studies into three codes: those published in peer-reviewed journals, institute reports, and unpublished working papers. 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. We also included one new working paper that was still undergoing peer review at the time of our review.

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 means of providing a comparable measure of sample size across studies. Importantly, the level of randomization and analyses varied across studies making this measure an imperfect proxy for the statistical power of a given study.

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 our sample into studies of coaching that were content-specific versus those that focused on more general pedagogical practices (e.g., programs that focused on improving students' social and emotional skills, including their behavior in class). 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.

Complementary Treatment Elements: Many of the studies included in our 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. Through an interactive process, we found that these three codes captured all additional and complementary resources that teachers received.

In-person versus Virtual Coaching: We coded coaching models as either delivered in person or virtually. In one instance where coaching was delivered as a combination of both we coded the model as in-person coaching given that even one-time in-person meetings may be central to establishing productive relationships (Powell, Diamond, Burchinal, & Koehler, 2010).

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 othehr 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.

Meta-Analytic Methods

We arrive at our pooled effect sizes using meta-analytic methods that produce precision weighted estimates and account for the clustered nature of our data (Hedges, Tipton, & Johnson, 2010; Tanner-Smith, Tipton, & Polanin, 2016). Our inclusion criteria and coding process produced a total of 142 effect sizes for instructional outcomes and 79 effect sizes for achievement outcomes across the 37 studies. Many studies contributed more than one effect size for a given outcome type because multiple measures of a given type were used (e.g., studies that reported dimension-level scores from an observation instrument of teachers' classroom practice rather than summative scores), or measures of the same type were captured at multiple points in time.

Our broad approach was intended to include as many treatment effects as possible and to account for the clustering of effect size estimates within studies through the analytic approach that we describe below. However, when there were multiple treatment-control contrasts from the same study, we only included effect sizes from contrasts that most closely aligned with those from other studies. For example, in a study by Garet et al. (2008), study participants were randomized to three groups (PD workshop, coaching plus PD workshop, and business as usual), which allowed for multiple comparisons. Here we focused on the coaching plus PD workshop versus business as usual contrast as this was the design taken by most other studies without multiple contrasts.

We estimate a standard random effects meta-analytic model where our 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 stocastic error term. The variance of the study-level random effects, $Var(u_j) = \tau^2$, is the between-study variance component.

We examine the association between components of different coaching models and our 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 both of these models using Robust Variance Estimation (RVE) methods (Hedges et al., 2010; Tanner-Smith et al., 2016) which account for the non-independence of effect sizes within studies through a method that is analogous to clustered standard errors. As with standard meta-analytic methods, RVE applies approximately inverse variance weights to improve the precision of our estimates (Hedges et al., 2010). However, RVE weights account for both the differing degrees of precision across studies (due to differences in sample sizes, level of randomization, predictive power of covariates, etc.) as well as the dependence of effect size estimates from the same study. 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 (represented by the ".") for the n_j effect sizes in study j for outcome k, and τ^2 is the estimated between-study variance component estimated via methods of moments. As equation 3 shows – effect sizes that are estimated from larger studies with greater precision are given larger weights while effect sizes from studies that contribute multiple effect sizes estimates are given less weight.

Findings

Characteristics of Included Studies

Our search yielded a total of 37 studies that met our inclusion criteria. We present descriptive statistics on these studies in Table 1. The vast majority of studies were published in peer-reviewed journals (n=30) and used experimental research designs (n=31). Notably, every single causal study of teacher coaching we identified was published on or after 2008, suggesting that researchers were immediately responsive to the findings of Yoon et al. (2007) that highlighted the lack of a causal research base on PD programs. Among these studies, 26 evaluated content-specific coaching programs while 11 assessed coaching programs for general 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=22 for reading, compared to n=2 for math and n=2 for science). Twenty-nine of the 37 studies included teachers who worked in early childhood education or elementary schools, another consequence of the early support for literacy coaching programs.

Across the studies we examined, 89% 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 early childhood 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. Twelve of the 37 studies also provided teachers with instructional content materials such as curriculum, lesson plans, and guide books. Another 11 studies relied heavily on video as a coaching resource: seven studies where teachers met virtually with a coach to discuss video recordings of their classroom instruction and had access to videos of other teachers implementing a range of instructional techniques with their classes, two studies where the program only used videos for virtual coaching, and another two where the program only provided video libraries.

Among the studies in our analytic sample, we found that the reported number of hours teachers worked one-on-one with a coach varied widely across coaching programs. Six studies reported coaching dosages of ten hours or less while six studies reported 30 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 six 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 our meta-analysis.

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 .14 SD and .92 SD. Effects on achievement also are distributed approximately normally with a positive skew and an interquartile range between .01 SD and .21 SD.

Turning to results from our meta-analytic models, we find large positive effects of coaching on teachers' instructional practice (see Table 2, Column 1). We find a pooled effect size of .57 standard deviations (SD) across all 25 studies that included a measure of instructional practice as an outcome. Estimates for coaching models focused on improving teachers' general practices are slightly larger (.71 SD) than estimates for content-specific coaching programs (.51 SD), although these estimates are not statistically different from each other.

We find that, on average, teacher coaching also has a meaningful positive impact on student achievement (see Table 2, Columns 2-5). Across all coaching models, we estimate that coaching raised student performance on standardized tests by .11 SD based on effect sizes reported in 21 studies. This estimate pools achievement tests across reading, math, and science in order to provide a broad picture of coaching effectiveness. We see similar patterns of meaningful effects on pooled student achievement when we disaggregate our results by program type: content-specific versus general (see Table 2 Rows 2-3). These sub-group analyses are somewhat limited given only three of nine studies that evaluated general coaching programs examined effects on student achievement. 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.

When we disaggregate results by the content area of the student assessment (see Table 2 Columns 3-5), we see that the vast majority of effect sizes on achievement are from reading assessments. This finding makes sense given that many of the coaching programs focused specifically on building teachers' literacy skills or were focused on broad instructional practices in early childhood/early elementary grades in which literacy is a key focus of instruction. Our pooled estimate of the effect of all coaching models on students' reading achievement are nearly identical (.12 SD) to results described above. In Appendx Table A1, we further disaggregate results by the content focus of the coaching program, in addition to the content focus of the student assessment. Here, we find a slightly larger pooled effect size estimate of literacy coaching programs on students' reading achievement of .14 SD across 14 studies.

For completeness, we also report subject-specific achievement results for math and science (see Table 2 Columns 4 and 5 for the mean effect size of all coaching programs on either math or science assessments, Appendix Table A1 Column 3 for the effect size of math-focused coaching programs on students' math achievement, and Appendix Table A1 Column 4 for the effect size of science-focused coaching programs on students' science achievement). However, we caution readers from making inferences based on these effect sizes estimates. These results have very limited generalizability and their associated standard errors are biased downward given that the RVE method does not fully correct for type I error in extremely small samples (Tipton & Pustejovsky, 2015).

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 our analytic sample to studies that employ causal research designs, these meta-analysis 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 key limitations on statistical power that prevent us from ruling out small but meaningful relationships in many cases. First, coaching model features 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 our data (Tanner-Smith et al., 2016). Given this, we report our 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 both instructional or achievement outcomes when coaching is combined with additional PD features, or when it is delivered in person versus virtually.

One exception is our 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 our 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 stand in contrast to previous findings on the importance of dosage in PD programs more broadly (Yoon et al., 2007) and 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; Hanssen, 2006; Scher & O'Reilly, 2009; Weiss & Miller, 2006). Our coded meta-analysis data afford a unique opportunity to examine this critical assumption empirically using experimental 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=9 of 37 total studies). [4] First, we averaged effect size estimates for each outcome within a study. Then, we conducted a weight-based analysis using the product of the average inverse variance of estimates for instructional and achievement outcomes.[5] 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

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

[5] Alternative weight schemes as well as unweighted analyses produce very similar results.

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 our 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 sort of analysis.

Across our analyses we find strong supporting evidence for the link between instruction and achievement. The correlation between effect sizes on instruction and achievement across these nine data points is .64. We illustrate the strength of this relationship in Figure 3. Interestingly, the magnitude of this relationship reveals that changes in student achievement appear to require relatively large improvements in instructional quality. Using a simple regression framework, we estimate that a one SD change in instruction is associated with a .15 SD change in achievement. This helps to explain why PD that results in only modest changes to teachers' instructional practice often does not lead to impacts on student achievement.

Sensitivity Analyses

We next examine the sensitivity of our estimates to the potential threat of missing data. Data may be missing from our analytic dataset due to publication bias or non-reported outcomes. These biases occur when studies that do not find statistically significant effects are not submitted or are 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 examine the sensitivity of our analyses by conducting a modified version of Duval and Tweedie's (2000) trim and fill method to account for the clustered nature of our data and the diverse range of coaching models in our 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 our centering estimate derived from our 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 our analytic sample. We also replicate this approach after collapsing our data to the study level by averaging effect sizes and variance estimates within studies by outcomes. As reported in Table 3, our adjusted estimates remain meaningfully large and statistically significant across both approaches. Pooled effect-size estimates are approximately .40 SD for instructional outcomes and .06 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

The results of our meta-analysis suggest that teacher coaching programs hold real promise for improving teachers' instructional practice and, in turn, students' academic achievement. These findings provide strong motivation to invest in efforts to scale up teacher coaching models, and to expand and improve upon the existing research base. Below we discuss how existing evidence can inform these efforts.

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. We test for evidence of potential scale-up implementation challenges by dividing our sample of studies into two groups following Wayne et al. (2008): *efficacy* trials (studies with samples of fewer than 100 teachers) and *effectiveness* trials (studies with samples of 100 teachers or more). Roughly speaking, studies included in our analyses 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 authors often playing a role in designing and delivering the coaching program to a small group of highly motivated volunteer teachers (e.g., Allen et al., 2015; Matsumara, Garnier, & Spybrook 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 in our sample grappled with recruiting and training a large coaching corps to deliver a more standardized program across a broad range of contexts where teachers had mixed levels of interest in the program (e.g., Gamse et al., 2008; 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 4, we estimate that smaller coaching programs improved classroom instruction by .78 SD and raised student achievement by .17 SD. These pooled effect sizes are approximately twice as large as the effects we find for larger coaching programs (.42 SD for instruction and .08 SD for achievement). The differences in effect size estimates across these subgroups are marginally significant for both instruction (*p*=.051) and achievement (*p*=.072) outcomes. 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. A recent study suggests one promising alternative is to pair classroom teachers in the same school with different strengths and weaknesses to support each other (Papay, Taylor, Tyler & Laski, 2016). Another approach taken by many districts has been to fold coaching into the observation component of new teacher evaluation systems. Both theory (Herman & Baker, 2009) and case-study analyses (Kraft & Gilmour, in press) 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 principals' and school leaders' existing responsibilities with little training or support is unlikely to result in intensive or sustained coaching.

Web-based virtual coaching offers one promising new approach for matching effective coaches with teachers. Leveraging video-based technology can lower coaching costs by eliminating commute time and increasing the number of teachers with whom an individual coach can work. It also has the potential to increase access to high-quality coaches for schools or districts without local expertise and to reduce possible reservations among teachers about mixing PD and evaluator roles by having their coach be both physically separate from and unaffiliated with their school. The small and insignificant point estimates for virtual coaching in our exploratory meta-regression models suggest this approach may 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 on-site 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 our review likely benefit from the non-random sample of teachers and schools that volunteered to participate in the studies. The two studies in our sample in which individual teachers or schools did not volunteer to participate but instead were part of district or state-wide rollouts provide suggestive evidence of 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 small, between .01 SD and .03 SD. A federally funded evaluation of Reading First which involved over 240 coaches across 13 states found moderately sized effects on teachers' instruction across first, second, and third grade classrooms (from .05 SD to .46 SD) and no significant effects on students' performance on state reading tests. One exception was significant effects on a supplemental test of decoding skills in first grade. In these two studies, it is not possible to disentangle whether 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, these studies point to 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 a 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 and an effort 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, in press). This suggests that building environments where providing and receiving constructive feedback is a regular part of teachers' professional work may be key condition for the success of scale-up efforts.

Realizing the results of coaching programs included in this meta-analysis at 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.

Directions for Future Research

Our 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. Our results also point to the relative lack of evidence on content-based coaching programs for subjects other than reading and literacy. New and emerging models of coaching such as real-time coaching where coaches stand in the back of a room and provide “bug-in ear” guidance to teachers via an earpiece also present fertile areas for future research (Ottley et al., 2016).

It also will be important to examine more closely which specific instructional practices are affected by coaching and what 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, an emerging body of research indicates that 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 (Blazar & Kraft, 2015). In turn, different teaching skills have differential impacts on a range of student outcomes (e.g., academic achievement, behavior, self-efficacy; Blazar & Kraft, 2016). 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 United States.

Similarly, we see a need for studies to move beyond efficacy trials of coaching models 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 on-site versus virtual coaching would be particularly informative in these efforts. 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 researchers are not primarily responsible for program implementation. Building the knowledge base about how to scale up coaching programs is, in our view, the single most important area for future research if coaching is going to have a meaningful impact on instructional quality.

Finally, all futures studies would benefit from examining outcomes in the year after the coaching program ends. It may be that sustained improvements in some areas of teaching practice require more than one year of coaching, even if at a reduced level. Among the 37 studies we reviewed, only four reported outcomes from a follow-up year (Allen et al., 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 who improve their practice by participating in coaching continue to implement the new practices they have learned 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 a 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.

How to Improve Experimental Studies on Teacher Coaching

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

- The features and approach of the coaching program
- 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 and characteristics of coaches as well as any training and support they receive
- 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 not only to inform the research design process itself but also will provide essential information to researchers and practitioners interested in replicating or adopting these models.

From a design and analysis standpoint, 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 (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, studies must examine the threat posed by attrition from their sample, which is common across most studies of PD. Attending to these design and analysis features will help to strengthen the quality of the future research literature.

Conclusion

By pooling results from across 37 causal studies of teacher coaching, we find large effects of coaching on both instruction and achievement. These positive effects compare favorably when contrasted with the larger body of literature on teacher PD (Yoon et al., 2007). Further, pooled effects on standardized achievement in reading of .12 SD are larger than pooled estimates from causal studies of almost all other school-based interventions reviewed by Fryer (2016) including student incentives, teacher pre-service training, merit-based pay, general PD, data-driven instruction, extended learning time, school choice, and charter schools. Only high-dosage tutoring (e.g., Blachman et al., 2004; Kraft, 2015) and school-level comprehensive reform models such as Success for All (Borman et al., 2007) and Reading First (Schwartz, 2005) were found to be more effective, but with far fewer studies contributing to these pooled effects.

Comparisons of effect sizes across school-based interventions are informative, but from a policy perspective these effects 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. Unfortunately, the existing literature lacks the necessary information about program costs to conduct a reliable cost benefit analysis. As researchers and practitioners continue to innovate, they likely will discover ways to reduce costs while maintaining the efficacy of coaching. We highlight some of these possibilities, including virtual coaching, above. 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 a meaningful financial investment. 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.

The growing literature on teacher coaching provides a much needed evidentiary base for future directions in teacher development policy and practice. The vast size of the teacher workforce in the United States requires development efforts that help teachers in classrooms across the country improve their skills. Teacher coaching models provide a flexible blueprint for these efforts. Thier ultimate success will depend on the administrators, coaches, and teachers doing this work on the ground. As a research community, we must help them grapple with the challenges of implementing coaching at scale in a sustainable and 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

*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.

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

*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/toc/esj/current

*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.

*Blazar, D., & Kraft, M. (2015). Exploring mechanisms of effective teacher coaching: A tale of two cohorts rom a randomized experiment. *Educational Evaluation and Policy Analysis, 37*(4), 542-566. doi: 10.3102/0162373715579487

Blazar, D., & Kraft, M. (online 2016). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis.*

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2015). *Attending to general and content-specific dimensions of teaching: Exploring factors across two observation instruments*. Working Paper. Cambridge, MA: National Center for Teacher Effectiveness. Retrieved from http://scholar.harvard.edu/files/david_blazar/files/blazar_et_al_attending_to_general_and_content_specific_dimensions_of_teaching.pdf

*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*.

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal*, *44*(3), 701-731.

Bryk, Anthony, and Barbara Schneider. *Trust in schools: A core resource for improvement*. Russell Sage Foundation, 2002.

*Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*(3), 430-454. Retrieved from http://www.journals.uchicago.edu/toc/esj/current

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.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. Teachers College Record, 102(2), 294–343.

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

*Conroy, M. A., Sutherland, K. S., Algina, J. J., Wilson, R. E., Martinez, J. R., & Whalon, K. J. (2014). Measuring teacher implementation of the BEST in CLASS intervention program and corollary child outcomes. *Journal of Emotional and Behavioral Disorders*, 1-12. doi: 10.1177/1063426614532949

Cook, T. D. (2001). Sciencephobia. *Education Next*, *1*(3).

Cornett, J., & Knight, J. (2009). Research on coaching. Coaching: Approaches and perspectives, 192-216.

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

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.

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.

*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*. Sage.

Fryer Jr, R. G. (2016). *The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments* (No. w22130). National Bureau of Economic Research.

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.

^*Gamse, B.C., Jacob, R.T., Horst, M., Boulay, B., & Unlu, F. (2008). Reading First Impact Study Final Report (NCEE 2009-4038). Washington, DC: *National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.* Retrieved from http://ncee.ed.gov

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.

^*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.* Retrieved from http://ies.ed.gov/ncee

^*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. Retrieved from http://ncee.ed.gov

Garet, M. S., Heppen, J.B., Walters, K., Parkinson, J., Smith, T.M., ..., Wei, T.E. (2016, November). 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* (No. 691d9603eb074051b57684e4affae4d4). Mathematica Policy Research.

*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.

Hanssen, C. E. (2006, October). The Milwaukee Mathematics Partnership: A pathmodel for evaluating teacher and student effects. Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466-479.

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

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.

Hill, H. C. (2007). Learning in the teacher workforce. *Future of Children, 17*(1), 111-127.

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

Hill, H. C., Blazar, D., & Lynch, K. (2015). Resources for Teaching. *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, and 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*. 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.

Jackson, C. K. (2016). *What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes* (No. w22226). 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.

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

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*. ASCD.

Kraft, M.A. (2015). How to make additional time matter: Extending the school day for individualized tutorials. *Education Finance and Policy. 10(1)*, 81-116.

Kraft, M.A. & Blazar, D. (in press). Individualized coaching to improve teacher practice across grades and subjects: New experimental evidence. *Educational Policy.*

Kraft, M.A. & Gilmour, A. (in press) Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly*.

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.

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*.

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

*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. Access at http://www.ncl.ac.uk/cflat/news/documents/CoachingSkillsTWFinalwebPDFv3.pdf

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. Retrieved from http://www.journals.uchicago.edu/toc/esj/current

*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

*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.

*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.

Murnane, R., & Willett, J. (2011). *Methods matter. Improving causal inference in educational and social science research*. 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. Retrieved from http://www.journals.uchicago.edu/toc/esj/current

^*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.

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

Ottley, J. R., Coogle, C. G., Rahn, N. L., & Spear, C. F. (2016). Impact of Bug-in-Ear Professional Development on Early Childhood Co-Teachers' Use of Communication Strategies. *Topics in Early Childhood Special Education*, 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.

^*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.air.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. Retrieved from http://www.journals.elsevier.com/early-childhood-research-quarterly/

*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.

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. *National Center for Education Evaluation and Regional Assistance*.

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

Richard, A. 2003. 'Making our own road': The emergence of school-based staff developers in America's public schools. New York: 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). *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/hel-home/issues/20_4/helarticle/school-based-coaching_269

^*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. Retrieved from http://www.journals.uchicago.edu/toc/esj/current

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

*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. Retrieved from http://www.journals.elsevier.com/teaching-and-teacher-education

*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.

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

Schwartz, R. M. (2005). Literacy Learning of At-Risk First-Grade Students in the Reading Recovery Early Intervention. *Journal of Educational Psychology*, *97*(2), 257.

Showers, B. (1984). Peer Coaching: A Strategy for Facilitating Transfer of Training. A CEPM R&D Report.

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.

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.

*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

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 1076998615606099.

*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/teaching-and-teacher-education

^*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

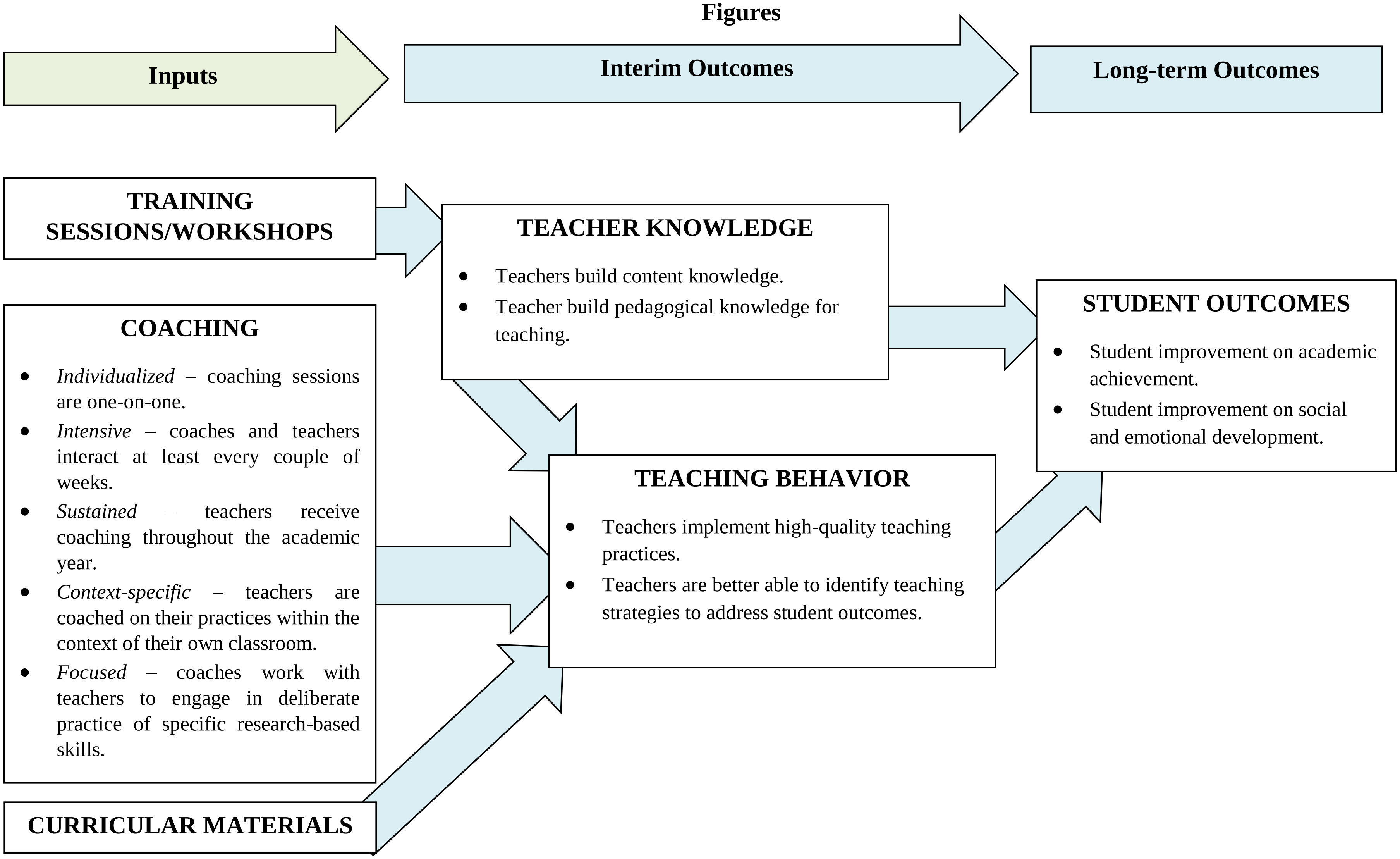
Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational researcher*, *37*(8), 469-479.

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.

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.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

*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

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

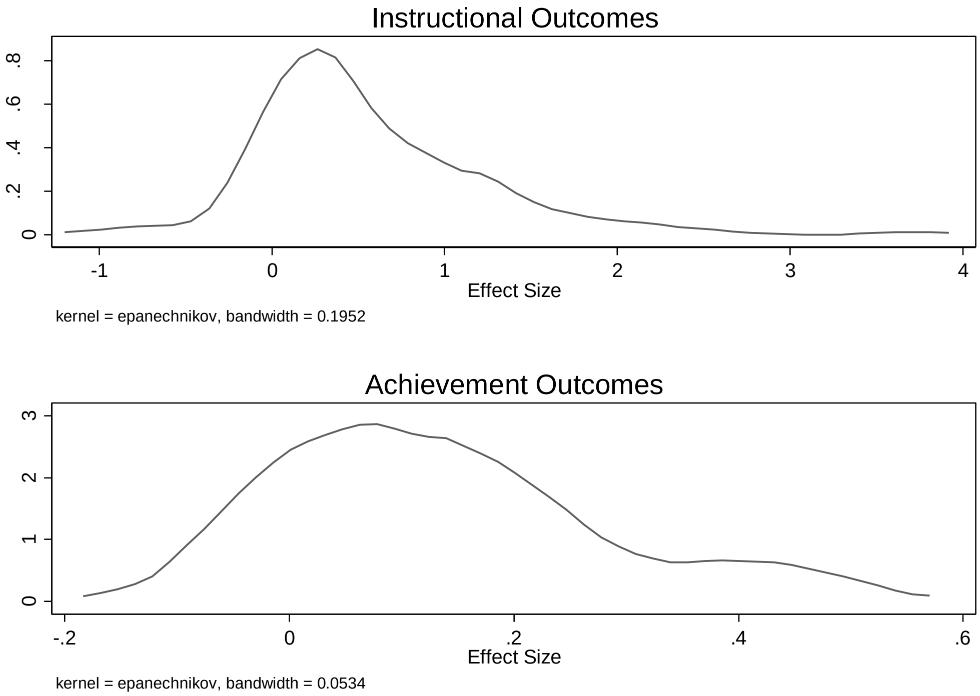


Figure 2: Kernel density plots of effect sizes for instructional and achievement outcomes.

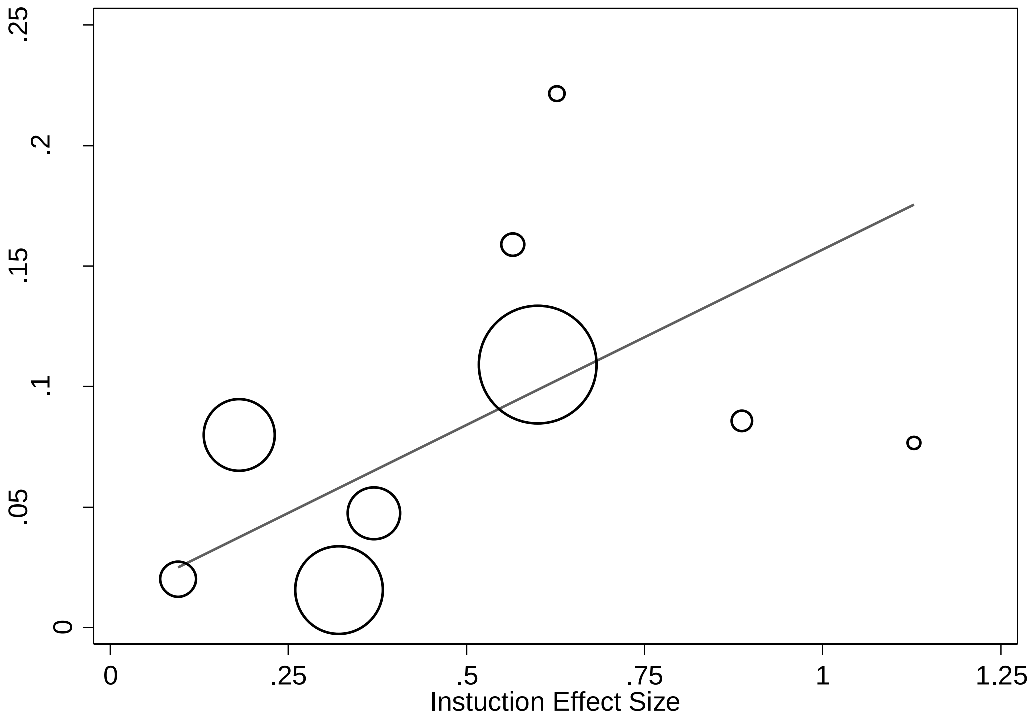


Figure 3. The relationship between coaching program effects on instruction and achievement.

Note: 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.

Tables

Table 1. Study Characteristics

	Count	Percent
Source		
Institute Report	6	0.16
Peer-reviewed Journal	30	0.81
Unpublished Working Paper	1	0.03
Year of Publication		
2008	4	0.11
2009	2	0.05
2010	8	0.22
2011	8	0.22
2012	1	0.03
2013	1	0.03
2014	7	0.19
2015	5	0.14
2016	1	0.03
Research Design		
Randomized Control Trials	32	0.86
Quasi-experiment	5	0.14
Level of Randomization (n=31)		
Teacher	14	0.45
School	15	0.48
District	2	0.06
Teacher Sample Size		
50 or less	8	0.22
51 to 100	11	0.30
101 to 150	7	0.19
151 to 300	6	0.16
300 plus	5	0.14
Not reported	2	0.05
Coaching Model Type		
Content-Specific	26	0.70
Math	2	0.05
Reading	22	0.59
Science	2	0.05
General Practices	11	0.30
School Level		
Pre-K	15	0.41
Elementary	14	0.38
Middle	15	0.41
High	5	0.14
Complementary Treatment Elements		
Any Complementary Treatment	33	0.89
Group Trainings	30	0.81
Instructional Content	12	0.32
Video Library	9	0.24

Delivered in Person		
Yes	28	0.76
No	9	0.24
Coaching Dosage (# of hours of one-on-one coaching)		
10 or less	6	0.24
11 to 20	8	0.32
21 to 30	5	0.20
30 or more	6	0.24
Not reported	12	
Total PD Dosage (# of hours)		
20 or less	8	0.32
21 to 40	6	0.24
41 to 60	8	0.32
60 or more	6	0.24
Not reported	9	
n(k)	37(221)	

Table 2. Pooled Effect Size Estimates

	Classroom Observations	Achievement (Pooled)	Reading Achievement	Math Achievement	Science Achievement
All Studies	0.567***	0.112***	0.124**	0.022	0.111
	(0.076)	(0.025)	(0.032)	(0.044)	(0.025)
n(k)	142(25)	79(21)	57(17)	19(4)	3(2)
Content-Specific	0.507***	0.120***	0.136**		
	(0.071)	(0.024)	(0.031)		
n(k)	90(16)	74(18)	53(14)		
General Practices	0.705**	0.096	0.101		
	(0.198)	(0.139)	(0.126)		
n(k)	52(9)	5(3)	4(3)		

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

Table 3. 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	0.567***	0.112***	.518***	.107***
	(0.076)	(0.025)	(0.065)	(0.025)
n(k)	142(25)	79(21)	25	21
	Panel B: Estimates with Imputed Missing Studies			
All studies	0.396**	0.064*	.396***	0.066*
	(0.108)	(0.028)	(0.079)	(0.028)
n(k)	169(30)	99(27)	33	28

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

Table 4. Pooled Effect Size Estimates

	Classroom Observations	Achievement (Pooled)
All Studies	0.567***	0.112***
	(0.076)	(0.025)
n(k)	142(25)	79(21)
Efficacy Trials (n Teachers <100)	0.783***	0.166**
	(0.151)	(0.032)
n(k)	67(13)	27(10)
Effectiveness Trials (n Teachers ≥100)	0.415***	0.077*
	(0.068)	(0.033)
n(k)	75(12)	52(11)

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

Appendix

Table A1. Pooled Effect Size Estimates by the Subject of Content-Focused Coaching

	Classroom Observations	Reading Achievement	Math Achievement	Science Achievement
Reading	0.507***	0.136**		
	(0.076)	(0.031)		
	84(14)	53(14)		
Math			0.077	
			(0.063)	
			14(2)	
Science				0.111
				(0.025)
				3(2)

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

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Group Training	0.177			0.171						
	(0.151)			(0.177)						
Instructional Content		0.053		0.048						
		(0.159)		(0.166)						
Video Library			-0.070	-0.035						
			(0.141)	(0.160)						
Total # Complementary PD Features					0.042					
					(0.107)					
Virtual Coaching						-0.100				
						(0.138)				
Coaching Dosage							-0.004			
							(0.006)			
11-20 Coaching Hours								0.381		
								(0.402)		
21-30 Coaching Hours								-0.069		
								(0.214)		
31 or More Coaching Hours								0.004		
								(0.210)		
Total PD Dosage									-0.002	
									(0.003)	
21-40 Total PD Hours										0.322
										(0.166)
41-60 Total PD Hours										0.235
										(0.391)
61 or More Total PD Hours										0.007
										(0.176)
Intercept	0.430***	0.555***	0.598***	0.435*	0.507*	0.605***	0.735***	0.580***	0.712***	0.492***
	(0.120)	(0.108)	(0.109)	(0.206)	(0.202)	(0.105)	(0.183)	(0.091)	0.172	(0.111)
n	142	142	142	142	142	142	107	107	113	113

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

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Group Training	0.038			0.042						
	(0.050)			(0.054)						
Instructional Content		-0.020		-0.037						
		(0.040)		(0.056)						
Video Library			0.007	0.025						
			(0.055)	(0.074)						
Total # Complementary PD Features					0.005					
					(0.021)					
Virtual Coaching						0.030				
						(0.066)				
Coaching Dosage							0.000			
							(0.001)			
11-20 Coaching Hours								-0.117		
								(0.115)		
21-30 Coaching Hours								-0.062		
								(0.107)		
31 or More Coaching Hours								-0.103		
								(0.112)		
Total PD Dosage									-0.001	
									(0.001)	
21-40 Total PD Hours										0.102
										(0.077)
41-60 Total PD Hours										-0.027
										(0.095)
61 or More Total PD Hours										-0.027
										(0.072)
Intercept	0.085*	0.119***	0.111***	0.089*	0.107*	0.106***	0.105*	0.184	0.147**	0.121
	(0.039)	(0.036)	(0.030)	(0.044)	(0.044)	(0.029)	(0.047)	(0.106)	0.053	(0.062)
N	79	79	79	79	79	79	49	49	53	53

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