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

Matthew A. Kraft
Brown University

David Blazar
Harvard University

Dylan Hogan
Brown University

November 2016

Updated: December 2017

Abstract

Teacher coaching has emerged as a promising alternative to traditional models of professional development. We review the empirical literature on teacher coaching and conduct meta-analyses to estimate the mean effect of coaching programs on teachers' instructional practice and students' academic achievement. Combining results across 60 studies that employ causal research designs, we find pooled effect sizes of 0.49 standard deviations (SD) on instruction and 0.18 SD on achievement. Much of this evidence comes from literacy coaching programs for pre-kindergarten and elementary school teachers. Although these findings affirm the potential of coaching as a development tool, further analyses illustrate the challenges of taking coaching programs to scale while maintaining effectiveness. Average effects from effectiveness trials of larger programs are only a fraction of the effects found in efficacy trials of smaller programs. We conclude by discussing ways to address scale-up implementation challenges and providing guidance for future causal studies.

Suggested Citation:
Kraft, M.A., Blazar, D., Hogan, D. (in press). The effect of teaching coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*.

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

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

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

In the U.S. and elsewhere, 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 (Garet et al., 2008; Garet et al., 2011; Garet et al., 2016; Glazerman et al., 2010; Harris & Sass, 2011; Jacob & Lefgren, 2004; Randel et al., 2011). These findings are particularly troubling given the lasting impact teachers have on individual students' long-term outcomes and on the economy as whole (Chetty, Friedman, & Rockoff, 2014; Hanushek, 2011; Jackson, 2016). The need for further training only has grown in recent years as professional expectations for teachers continue to rise and states adopt new "college- and career-

[1] Arriving at an exact estimate of total expenditures on PD is complicated by the fact that U.S. federal requirements have districts report expenditures on PD as part of an "Instructional staff services" category which also includes expenditures for curriculum development, libraries, and media and computer centers. Most studies find that districts allocate 3 to 5 percent 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).

ready" standards that require teachers to integrate higher-order thinking and social-emotional learning into the curriculum.

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

We extend this work by reviewing the causal evidence on one specific PD model that is centered on several of these "critical features" and that has gained increasing attention in recent years: teacher coaching. We define coaching programs broadly as all in-service PD programs where coaches or peers observe teachers' instruction and provide feedback to help them improve. While coaching fits under the broader umbrella of PD and teacher learning, we see it as distinct from most program offerings, which still consist of short-term and generalized workshops

(Darling-Hammond et al., 2009). In contrast, coaching is intended to be individualized, time-intensive, sustained over the course of a semester or year, context-specific, and focused on discrete skills.

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. bBeginning in the late 1990's in the U.S., federal legislation aimed at strengthening the quality of reading instruction helped formalize and fund coach positions for reading teachers in schools (Denton & Hasbrouck, 2009). These included the passage of the Reading Excellence Act in 1999, No Child Left Behind (NCLB) in 2002, and the reauthorization of the Individuals with Disabilities Education Act (IDEA) in 2004. The legacy of these investments is evident today in the wide range of established literacy coaching programs and the preponderance of research focused on literacy coaching models.

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

complement these works by conducting the first meta-analysis of studies examining the causal effect of teacher coaching on instructional practice and student achievement.

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

We identified 60 studies of teacher coaching programs in the U.S. and other developed countries that both used a causal research design and examined effects on instruction or achievement.[2] This new body of causal research on teacher coaching suggests that IES funding and scholars advocating for wider use of causal methods were successful at pushing the field in this direction. We focus our review of the coaching literature on the U.S. and other developed nations because the vast majority of the theoretical and empirical research comes from these settings. Although there is an emerging body of causal research on PD and coaching in developing nations, our approach allows us to define a clear population of interest and to avoid generalizing across programs implemented in substantially different contexts.[3] Further, research

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

[3] We identified four causal studies on coaching in developing contexts: Albornoz et al. (nd), Harvey (1999), Piper & Zuilkowski (2015), and Sailors et al. (2014). For a synthesis of the evidence on in-service teacher training programs in the international context see Timperley et al. (2008) and Popova, Evans and Arancibia (2016).

on other teacher-oriented programs such as financial incentives suggest that outcomes may differ substantially across developing and developed countries (Ganimiam & Murnane, 2016; Gneezy, Meier, & Rey-Biel, 2011); yet we would be underpowered to test formally for such differences given the small number of causal studies on teacher coaching available from developing countries.

The use of meta-analytic methods to analyze these studies affords the ability to explore questions about teacher coaching that no single experimental trial can address. First, we are able to better understand the efficacy of coaching as a general class of PD by analyzing results across a range of coaching models. Second, the large financial and logistical costs of conducting experimental evaluations of teacher coaching programs has resulted in many individual studies that are underpowered. Meta-analysis techniques leverage the increased statistical power afforded by pooling results across multiple studies. This is critical for determining whether common findings of positive effect sizes that are not statistically significant are due to limited statistical precision or chance sampling differences. Third, meta-analytic regression methods facilitate a comparison of different coaching models and a closer examination of specific design features that may drive program effects, such as the size of coaching programs, pairing coaching with other PD elements, in-person versus virtual coaching, and coaching dosage. To date, questions of the effectiveness of individual design features have been explored by only a handful of studies (e.g. Blazar & Kraft, 2015; Marsh et al., 2008; Ramey et al., 2011).

Our analyses are driven by three primary research questions:

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

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

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

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

Method

Working Definition of Teacher Coaching Interventions

Although the majority of teacher coaching programs share several key program features, no one set of features defines all coaching models. At its core, "coaching is characterized by an observation and feedback cycle in an ongoing instructional or clinical situation" (Joyce & Showers, 1981, p.170). Coaches are thought to be experts in their field who model research-based practices and work with teachers to incorporate these practices into their own classrooms (Sailors & Shanklin, 2010). However, in our review of the literature we encountered multiple, sometimes conflicting, working definitions of teacher coaching. Some envision coaching as a form of implementation support to ensure that new teaching practices – often taught in an initial training session – are executed with fidelity (Devine et al., 2013; Kretlow & Bartholomew, 2010). Others see coaching as a direct development tool that enables teachers to see "how and why certain strategies will make a difference for their students" (Russo, 2004, p. 1; see also Richard, 2003). Still others describe multiple types of coaching, each with their own objectives. For example, "responsive" coaching aims to help teachers reflect on their practice, while

"directive" coaching is oriented around the direct feedback coaches provide to strengthen teachers' instructional practices (Ippolito, 2010). In line with these multiple perspectives, Gallucci et al. (2010) describe coaching as "inherently multifaceted and ambiguous" (p. 922). Coaches often take on these roles and others, including identifying appropriate interventions for teacher learning, gathering data in classrooms, and leading whole-school reform efforts.

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

Because improvements in teacher skill and classroom practice cannot be divorced from improvements in teacher knowledge (Hill, Blazar, & Lynch, 2015c), coaching rarely is

implemented on its own. Often, coaching is combined with training sessions or courses in which teachers are taught new skills or content knowledge (Kretlow & Bartholomew, 2010). It also may be used to develop teachers' abilities to work with new curricular materials or instructional resources. In a review of the literature on PD in early childhood settings, Schachter (2015) found that 39 of the 42 programs that included coaching as one element combined it with some other form of training (e.g., a workshop or course), and many also included additional resources such as curriculum materials or websites with video libraries.

We define coaching programs broadly as all in-service PD programs that incorporate coaching as a key feature of the model. The role of the coach may be performed by a range of personnel including administrators, master teachers, curriculum designers, external experts, and classroom teachers. 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 broad spectrum of models in this analysis that range from those focused on supporting the implementation of curriculum or pedagogical frameworks to those where the coaching process itself is the core development tool.

For the purposes of this review, we narrow this definition in two ways that we see as consistent with the broader literature on coaching programs. First, we exclude teacher preparation and school-based teacher induction programs. While these types of teacher training

are increasingly integrating observation and feedback cycles with instructional experts into their designs, it is difficult to disentangle coaching practices from the range of supports provided to new teachers as part of comprehensive induction programs (e.g., Glazerman et al., 2010). The role and goals of a mentor often are quite distinct from those of a coach. For example, mentors may provide advice on work-life balance and how to interact with school leadership, both of which are situated outside of teachers' classrooms. Second, we exclude coaching programs where coaches also provided direct services to students (e.g., Raver et al., 2009), given that it would be difficult to determine if any effects on student achievement were due to improvements in teachers' instruction or to these direct services.

Literature Search Procedures

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

Inclusion Criteria

We restricted the sample of studies published during or before 2017 using four primary

criteria pertaining to the sample, the intervention, the research design, and the outcomes.[4] First, we required that studies evaluate a PD program that incorporated teacher coaching as defined by our working definition above. Second, we limited this review to include studies where the sample was comprised of early childhood to 12th grade in-service teachers in the U.S. or other developed nations. Third, we required that studies employed an experimental or quasi-experimental research design capable of supporting causal inferences (Murnane & Willett, 2011; Shadish, Cook, & Campbell, 2002). We judged quasi-experimental designs as meeting this standard if they employed a regression discontinuity (no qualifying studies found), an instrumental variables approach with a justifiable instrument (no qualifying studies found), or a difference-in-differences design (e.g., Biancarosa, Bryk, & Dexter, 2010; Lockwood, McCombs, & Marsh, 2010; Teemant, 2014; Vogt & Rogalla, 2009). We excluded studies that relied principally on covariate adjustment without random assignment or used a pre-post design only for treated units given concerns that these strategies cannot adequately account for non-random selection. Fourth, we required that studies include at least one measure of teachers' 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. As causal research on teacher coaching continues to accumulate, meta-analytic work may examine effects on other important outcomes such as teacher knowledge and students' social-emotional competencies. In

[4] When multiple papers were published using the same set of data, we included papers when they reported results from different outcomes (Rimm-Kaufman et al., 2014 and Abry, Rimm-Kaufman, Larsen, & Brewer, 2013), different cohorts (Kraft & Blazar, 2017, and Blazar & Kraft 2015), or different years (Matsumura, Garnier, & Spybrook, 2013 and Matsumara, Garnier, & Spybrook, 2012) but chose only one of the studies when the samples, outcomes and periods of measurement were overlapping (Vernon-Feagans, Kainz, Hedrick, Ginsberg & Amendum, 2013 instead of Amendum, Vernon-Feagans, & Ginsberg, 2011).

the next section, we describe additional constraints placed on how these outcome measures were captured.

Outcomes

Instruction. Following the conceptual framework developed by Cohen, Raudenbush, and Ball (2003), we viewed instruction not simply as how teachers deliver lessons but rather as the interaction of teachers, students, and content within the context of classroom and school environments. Thus, we included scores from classroom observation instruments that captured 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 these measures of instruction only to those that were collected by outside observers blind to treatment status.[5] 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 identifies several unique domains of teaching practice (Blazar, Braslow, Charalambous, & Hill, 2015a; Hamre et al., 2013), it was not feasible to examine these constructs separately in these analyses. Studies used several different observation instruments or coding schemes that aimed to capture different elements of teachers' instructional practice; these instruments tended to align with the goals of the specific coaching program or the grade level of the students in the classroom. Observation instruments included rubrics that are well-established in the research literature and widely used by districts (e.g., Classroom Assessment Scoring System [CLASS], Early Language and Literacy Classroom Observation [ELLCO]), as well as lesser-known

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

instruments that were developed by researchers or coaching programs (e.g., Blazar & Kraft, 2015; Sailors & Price, 2015; Teemant, 2014). Because studies provided varying levels of information about these instruments, we were limited in our ability to assess the degree of overlap among specific dimensions. Relatedly, without access to the primary data, it was not possible to assess the measurement properties of scores produced by each of these instruments. However, most studies either used validated scales (e.g., CLASS, ELLCO), or reported high reliabilities (e.g., 80 percent or higher inter-rater agreement rates, internal consistency reliability of 0.80 or higher).

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

Coding Procedures

We coded studies for information needed to convert treatment effects on instruction and

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

Source and year of publication. We categorized the source of studies into three categories: peer-reviewed journal articles, 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.

Country of Study. The country in which a study was conducted

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

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

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

School level. We created a set of four indicators for the level of schooling that was the

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

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

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

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

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

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

Teacher and Coach Characteristics. We also searched articles for information about teacher and coach characteristics but found that inconsistent reporting approaches and a lack of detail limited our ability to construct formal codes. For example, authors most often reported information on teachers' years of teaching experience, but varied widely on how they

reported this information (e.g., mean and standard deviation, percentages of teachers who fell into discrete experience bins, range). For coach characteristics, authors were even less consistent in what they reported. Some provided information on teaching experience, while others focused on the training provided to coaches.

Meta-Analytic Approach

We arrive at pooled effect sizes using meta-analytic methods that produce precision weighted estimates and account for the clustered nature of the data (Hedges, Tipton, & Johnson, 2010; Tanner-Smith, Tipton, & Polanin, 2016). Our inclusion criteria and coding process produced a total of 186 effect sizes for instructional outcomes and 113 effect sizes for achievement outcomes across the 60 studies. Many studies contributed more than one effect size for a given outcome type because multiple measures were used (e.g., studies that reported dimension-level scores from an observation instrument of teachers' classroom practice), or because measures of the same type were captured at multiple points in time. Some studies also included multiple effect sizes due to multiple treatment groups (e.g., PD workshop, coaching plus PD workshop, and business-as-usual control in Garet et al., 2008). In these instances, we focused only on the treatment-control contrast that most closely matched the designs of other studies: coaching (plus any complementary activities) versus business-as-usual control.

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

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

Here, y_{ij}^{k} captures a given effect size *i* for outcome type *k* in study *j* where models for different outcome types are fit separately. Alpha, α, captures the pooled effect size estimate for outcome

k, u_j is the study level random effect, and ε_{ij}^k is the mean-zero stochastic error term.

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

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

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

We estimate all models using Robust Variance Estimation (RVE) methods (Hedges et al., 2010; Tanner-Smith et al., 2016) which account for both the differing degrees of precision across studies as well as the non-independence of effect sizes within studies through a method that is analogous to clustered standard errors.[6] RVE methods up-weight effect sizes that are estimated with greater precision (due to differences in sample sizes, level of randomization, predictive power of covariates, etc.) and down-weight estimates from studies that contribute multiple effect size estimates. In several instances, research teams published multiple studies by analyzing different outcomes from the same research project in different articles. We test the sensitivity of our inferences by recoding all studies that use data from the same research project as a single study and find that our results are unchanged.

Results

Characteristics of Included Studies

We present descriptive statistics on the 60 studies that met our inclusion criteria in Table 1 and include the full list of studies and associated codes in Appendix Table A1. Every study we identified was published on or after 2006 with the vast majority of studies in peer-reviewed

[6] Weights are constructed such that $w_{ij}^k = \frac{1}{n_j^k(v_{.j}+\tau^2)}$ where $v_{.j}$ is the mean of the individual *i* variances for the n_j effect sizes in study *j* for outcome *k*, and τ^2 is the estimated between-study random effect variance component from equation (1) [i.e., $Var(u_j) = \tau^2$] estimated via methods of moments

journals (n = 51). Fifty-six of the 60 studies employed experimental research designs. Forty studies evaluated content-specific coaching programs while 20 assessed coaching programs for general instructional pedagogy. Given the history of U.S. 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 = 34 for reading, compared to n = 2 for math and n = 3 for science). Fifty-one of the 60 studies included teachers who worked in pre-kindergarten centers or elementary schools, another consequence of the early support for literacy coaching programs. Twelve of the studies evaluated virtual coaching models where teachers recorded themselves teaching and discussed their instruction on a web-based platform with a virtual coach. Of these 13 virtual coaching studies, 10 evaluated versions of the My Teaching Partner program developed by Robert Pianta and colleagues at the University of Virginia Center for Advanced Study of Teaching and Learning.

Across the studies we examined, 90 percent evaluated coaching models that were combined with at least one additional PD element. This finding is nearly identical with Schachter's (2015) review of the literature on PD for pre-kindergarten educators. Coaching was combined most frequently with group trainings in the form of summer workshops and team training sessions during the academic year where coaches might demonstrate lessons or instructional practices (48 of 60). Twenty-two of the 60 studies also provided teachers with instructional content materials such as curriculum, lesson plans, or guide books. Another 14 studies supplemented coaching with video exemplars of other teachers delivering high-quality instruction.

We found that the reported number of hours teachers worked one-on-one with a coach varied widely across coaching programs. Sixteen studies reported coaching dosages of ten hours

or less while 14 studies reported 21 hours or more. The total PD hours for participating teachers also varied across programs with 13 interventions consisting of 20 total hours or less and 10 interventions consisting of 60 total hours or more. This wide variation in the dosage of coaching and total PD hours illustrates the substantial differences in the coaching programs included in this meta-analysis.

Because average teaching experience was not reported in a consistent metric across studies, we do not include this information in Table 1. For those studies that did report mean years of teaching experience, the average was approximately 11 years. Some studies focused specifically on early career teachers (e.g. Blazar & Kraft, 2016, while others focused on more veteran teachers (e.g. Pianta, Mashburn, Downer, Hamre, & Justice, 2008; Teemant, 2014; Vernon-Feagans et al., 2013; Vogt & Rogalla, 2009).

Effects on Instruction and Achievement

Kernel density plots of effect sizes on teachers' instruction and students' achievement help provide visual evidence and intuition for our pooled estimates. As shown in Figure 2, the distribution of effect sizes of coaching on instruction is distributed approximately normally with a long right-hand side tail. The magnitude of effects varies considerably, with an interquartile range between 0.17 SD and 0.92 SD. Effects on achievement also are distributed approximately normally with a positive skew and an interquartile range between 0.03 SD and 0.24 SD.

Turning to our primary meta-analytic results for instruction in Table 2, Column 1, we find large positive effects of coaching on teachers' instructional practice. Across all 43 studies that included a measure of instructional practice as an outcome, we find a pooled effect size of 0.49 standard deviations (SD). The associated standard deviation of the estimated random effect – a measure of the variation in effect sizes across programs – is 0.33 SD suggesting there exists

substantial variability across programs. Disaggregating these results among content-specific coaching programs and those that focused on general pedagogical practices produces consistent estimates of 0.51 SD and 0.47 SD, respectively. The content-specific coaching programs covered several different areas: reading, mathematics, and science. However, only studies in reading had sufficient sample sizes to report disaggregated results, which also are quite similar (0.51 SD). In results available upon request, we find similar point estimates for effects on instruction when comparing studies in the U.S. to the five international studies in our sample (.50 SD vs. .42 SD).

On average, teacher coaching also has a positive effect on student achievement as shown in Table 2, Columns 2-5. Across all coaching models, we estimate that coaching raised student performance on standardized tests by 0.18 SD based on effect sizes reported in 31 studies that included measures of students' academic performance. The associated standard deviation of the estimated random effects is 0.18 SD, again suggesting effects differ substantially across programs. Many of the achievement measures included in these analyses were selected or designed by researchers to be closely aligned with the coaching programs. Ten studies provide the opportunity to evaluate the effect of coaching on state standardized tests, which are intended to assess broad domains of knowledge and skills. In supplemental analyses, we estimate a more moderate pooled effect on student achievement on state standardized tests of 0.12 SD (p=.04, k=31, n=10), although the associated 95% confidence interval includes 0.18 SD.

These overall effect size estimates pool achievement across reading, math, and science tests in order to provide a broad picture of coaching effectiveness. However, our ability to make general inferences about achievement gains across subjects is limited by the fact that three quarters of the total number of achievement effect sizes use reading assessments as the outcome

measure. Narrowing in on programs that target students' early reading skills, we find a nearly identical average treatment effect of 0.18 SD on improvements in this specific skill.[7]

We see smaller effects on student achievement for general coaching programs (0.07 SD, not significant) than content-specific programs (0.20 SD). This makes sense given that general coaching programs often are focused less on helping teachers improve students' test scores and more on developing teachers' abilities to support students' social and emotional development. This is also evident in the fact that only four of the 20 studies that evaluated general coaching programs examined effects on student achievement. However, due to small sample sizes for achievement effects of general coaching programs, we cannot statistically distinguish these estimates from each other (p=.24).

Next, we explore potential differences in coaching program effects across school levels by estimating effects for pre-kindergarten centers, elementary schools, middle schools and high schools separately. As shown in Table 3, no clear pattern emerges from these analyses. While treatment effects on student achievement appear larger for K-12 schools relative to pre-kindergarten programs, none of coefficients across schooling levels are statistically significantly different from each; this is true both for achievement and instructional outcomes. This suggests that coaching may be an equally effective intervention with teachers working at all school levels.

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, although we restrict the analytic sample to

[7] Pooled effects on math and science achievement are shown in Table 2 although neither estimate is statistically significant. Science outcomes are only available for science-specific coaching models. In a supplemental analysis where we focus only on math-specific coaching program, we find that the estimate for math achievement increases to 0.08 (p = .44, k = 14, n = 2).

studies that employ causal research designs, these meta-analytic regressions are descriptive in nature and do not capture the causal effect of a given program feature. Limited statistical power also prevents us from ruling out smaller relationships in many cases.

As shown in Table 4, we find that pairing coaching with group trainings is associated with 0.31 SD larger effect size on instruction and 0.12 SD larger effect size on achievement. Consistent with the theory of action outlined in Figure 1, this suggests that teachers may benefit from building baseline skills (e.g., content knowledge) prior to engaging directly with a coach. For instructional outcomes, pairing coaching with instructional resources and materials (e.g., curriculum) also is associated with greater gains (0.21 SD larger), while providing teachers with a video library is associated with more limited benefits (-0.27 SD smaller). We do not find any significant difference in effect sizes for coaching programs that were delivered in person or virtually, though our standard errors are too large to rule out even moderately sized differences.

Finally, for both measures of dosage – total hours of coaching, and total hours of PD when coaching is paired with other program features – we fail to find any evidence in support of the hypothesis that coaching must be high-dosage to be effective. We find very precisely estimate null effects for both instruction and achievement outcomes. In further analyses available upon request, we do not find any clear evidence of potential threshold effects or other non-linear functional forms when we model these relationships non-parametrically. These findings are generally consistent with Kennedy's (2016) graphical analysis of features of effective PD programs showing no consistent relationship between dosage and outcomes, but stand in contrast to previous findings on the importance of dosage in PD programs more broadly (Yoon et al., 2007). The lack of evidence supporting dosage effects suggests 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 PD models is that helping teachers improve the quality of their instructional practice will lead to improvements in student achievement (Cohen & Hill, 2000; Kennedy, 2016; Scher & O'Reilly, 2009; Weiss & Miller, 2006). Our coded meta-analysis data afford a unique opportunity to examine this critical assumption empirically using causal studies that estimate impacts on both instruction and achievement. Although we can interpret the effect of coaching on instruction and on achievement in a causal framework, we cannot do so for the relationship between instruction and achievement. Our theory of change posits that improvements in instruction cause student achievement to rise. However, it is also possible that coaching effects on achievement were mediated through avenues other than instructional improvement (e.g., preparation time out of class). As such, we view these analyses as exploratory in nature. Access to the original data from these studies would allow us to instrument for instructional measures via random assignment of coaching, and we encourage future studies to engage in this type of analysis.

We find supporting evidence for the link between instruction and achievement. Across a small sample of 20 studies from 16 research projects that included both outcome measures, the strength of the weighted correlation between averaged effect sizes on instruction and achievement is 0.37 (p = .16; see also Figure 3).[8] To arrive at this estimate, we averaged effect size estimates for each outcome within a research project. In addition to asking how effect sizes on instruction and achievement covary, we can interpret the magnitude of this relationship by examining how large of a change in achievement is associated with a given change in instruction. This analysis produces different results from the correlation above given that the combined set of effect-size estimates is not standardized (see Figure 2). Here, we find that changes in student

[8] All 20 studies used in constructing this sample are denoted with a ^ in the references.

achievement appear to require relatively large improvements in instructional quality. Using a weighted linear regression framework, we estimate that a 1 SD change in instruction is associated with a 0.21 SD change in achievement (p = .16).[9] This finding is consistent with a large body of literature documenting the weak relationship between educational inputs (instruction) and outputs (achievement) and helps to explain why PD that results in more modest changes in teachers' instruction often does not lead to impacts on student achievement.

Sensitivity Analyses

We examine the sensitivity of our estimates to three threats to internal validity: missing data, research design, and outliers. We begin by examining the degree to which our results may be a product of missing data caused by when studies that do not find statistically significant effects are not submitted or not accepted for publication, as well as when authors of published studies do not include the results of all available outcomes in a paper. We test the sensitivity of these findings by conducting a modified version of Duval and Tweedie's (2000) trim and fill method to account for the clustered nature of the data and the diverse range of coaching models in the analytic sample. Using this rank-based data augmentation technique, we estimate the number of missing effect sizes and impute these theoretically missing data points. This involves calculating the hypothetical data points needed to balance the spread of effect sizes across a centering estimate derived from the random effects model in equation 2. We do this first at the effect-size level by imposing a nested structure on the imputed data based on the average number of effect sizes per study in the analytic sample. We also replicate this approach after collapsing the data to the study level by averaging effect sizes and variance estimates within studies for a

[9] We weight correlation and regression estimates by the average sample size of all instructional and achievement effect sizes from a given research project. Weighting results using variance estimates from instructional effect sizes produces qualitatively similar results.

given outcome. As reported in Table 5, the adjusted estimates are attenuated, particularly for instructional outcomes, but remain statistically significant across both approaches. Pooled effect-size estimates are approximately 0.34 SD for instructional outcomes and 0.14 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.

A second area of possible concern focuses on the research design of included studies. The vast majority of studies are randomized control trials that are considered the gold standard of causal inference design (Murnane & Willett, 2011). Additional studies that met our inclusion criteria but used quasi-experimental designs all employed variants of difference-in-differences strategies that rest on two critical assumptions: parallel trends between treatment and comparison groups, and no simultaneous confounding of treatment effects (Murnane and Willett, 2011). Given limited information to assess these assumptions directly, we instead probe the sensitivity of our findings by restricting the sample to only include randomized control trials. Unsurprisingly, these results are quite similar to our main findings, with pooled effects of coaching on instruction of 0.45 SD and achievement of 0.18 SD (see Appendix Table A2).

Finally, given the large variation in effect sizes (see Figure 2), it is also possible that our results are driven by outliers. Visual inspection of the data as well as box and whisker plots suggest there exist few clear outliers in our data. Rather than make a subjective decision about what data points constitute outliers, we test the sensitivity of our results by removing the lowest and highest 5 percent of the effect sizes for each outcome. As shown in Appendix Table A3, our results are not driven by extreme values and remain largely unchanged after trimming the bottom and top 5 percent of estimates. We find pooled effects across all studies of 0.45 SD for instruction and 0.16 SD for achievement.

Discussion

In order to interpret the substantive significance of our findings, we consider several benchmarks described by Hill, Bloom, Black, and Lipsey (2008) and Lipsey et al. (2012): the observed effect of similar interventions, policy-relevant performance gaps, normative expectations for students' academic growth, and costs. Our estimates of the effect of coaching on teachers' instructional practice (0.49 SD) are larger than differences in measures of instructional quality between novice and veteran teachers' (0.2 to 0.4 SD; Blazar & Kraft, 2015). Effects on students' academic performance (0.18 SD) are of similar or larger magnitude than estimates of the degree to which teachers improve their ability to raise student achievement during the first five to ten years of their careers, with estimates ranging from 0.05 to 0.15 SD (Atteberry, Loeb, & Wykoff 2015; Papay & Kraft, 2015). Effects on achievement are also larger than pooled estimates from causal studies of almost all other school-based interventions reviewed by Fryer (2017) including student incentives, teacher pre-service training, merit-based pay, general PD, data-driven instruction, and extended learning time. Interventions of comparable effect sizes on achievement include comprehensive school reform (0.1 to 0.2 SD, depending on the school reform model; Borman, Hewes, Overman, & Brown, 2003), oversubscribed charter schools (0.04 SD to 0.08 SD per year of attendance; Chabrier, Cohodes, & Oreopoulos, 2016), large reductions in class size (roughly 0.2 SD; Krueger, 1999), high-dosage tutoring (0.15 to 0.25 SD; Blazar et al., 2015a; Blachman et al., 2004), and changes in curriculum (0.05 to 0.3 SD depending on the grade level and curriculum under investigation; Agodini et al., 2009; Koedel, Li, Springer, & Tan, 2017).

From a policy perspective, the effects of teacher coaching must be considered relative to program costs. Traditional on-site coaching programs are a resource-intensive intervention

simply due to the high personnel costs of staffing a skilled coaching corps. One cost analysis of coaching across three schools found per-teacher costs ranged from $3,300 to upwards of $5,200 (Knight, 2012). Unfortunately, the existing literature lacks the necessary information about program costs to conduct a reliable cost-benefit or cost-effectiveness analysis. As researchers and practitioners continue to innovate, they should explore ways to minimize costs while maintaining the efficacy of coaching. We highlight some of these possibilities, including virtual coaching, in the remaining part of our discussion and conclusion. However, if an instructional expert working one-on-one with teachers in person over a sustained amount of time remains at the core of effective coaching models, then this approach will always require fairly sizeable financial and human capital investments. Given the billions of dollars U.S. districts and others around the world 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 utilized more effectively. One approach would be to allocate resources to high-cost but effective PD programs for teachers most in need of support, such as coaching, rather than to lower-cost but less-effective programs for all teachers.

Taking Teacher Coaching to Scale

Decades worth of researchers have documented the significant challenges of taking education programs and reform initiatives to scale (Honig, 2006). Given the fundamental importance of implementation quality, major questions still remain about the feasibility of expanding teacher coaching across schools and districts. For example, researchers found that when a literacy PD program was modified for scalability by reducing coaching frequency, using trained research assistants as coaches, and providing written rather than in-person feedback it had no effect (Cabell et al., 2011). We explore this question in our data by examining the

relationship between the scale of a coaching program and its effect size. We illustrate this relationship graphically in Figure 4 using teacher sample size as a simple proxy measure for program size. This figure depicts a scatterplot of the average effect size by deciles of teacher sample size with the linear relationship from an OLS regression overlaid on top. Graphs for both instruction (Panel A) and achievement (Panel B) depict a clear negative relationship between program size and program effects, consistent with a theory of diminishing effects as programs are taken to scale.

We more formally test for evidence of potential scale-up implementation challenges by dividing the sample of studies into two groups following Wayne et al. (2008): *efficacy* trials that examine small programs under conditions that are intended to be as conducive as possible to maximizing effects versus *effectiveness* trials that test larger-scale programs often implemented across a range of settings with more limited support. We approximate this distinction in our sample by comparing effects from studies with samples of fewer than 100 teachers to studies with more than 100 teachers. While this categorization approach is imperfect, it provides a simple and objective way to examine differences in outcomes between smaller versus larger programs. In the sample, the smaller-scale programs generally evaluated coaching programs with no more than 50 teachers and a handful of coaches (e.g., Allen et al., 2015; Matsumara et al.,2012; McCollum, Hemmeter, & Hsieh, 2013). These programs often were tailored specifically for teachers who were motivated to participate and the school contexts in which they work, suggesting that they were implemented under best-case conditions. In contrast, the larger programs with 100 or more teachers generally required recruiting and training a sizeable coaching corps to deliver a more standardized program across a broader range of contexts where teachers were more likely to have mixed levels of interest in the program (e.g., Garet et al., 2008,

2011; Lockwood et al., 2010).

Comparing pooled effect sizes estimates for efficacy versus effectiveness trails suggests that coaching can have an impact at scale but that scale-up implementation challenges likely attenuate this effect. As reported in Table 6, we estimate that smaller coaching programs improved classroom instruction by 0.63 SD and raised student achievement by 0.28 SD. These pooled effect sizes are approximately twice the size of effects on instruction for larger programs (0.34 SD) and three times the size of effects on achievement for larger programs (0.10 SD), with both differences statistically significant at the .05 level. Publication bias may explain some of this difference if efficacy trials of smaller programs are less likely to be published due to a lack of statistical significance. Many of the effectiveness trials of larger programs 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 smaller-scale demonstration models.

We next consider likely factors that contribute to the smaller effects of larger-scale coaching programs and ways that practitioners and policymakers might address them. 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 for smaller coaching programs. 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. However, this strategy comes with the tradeoff of potentially removing highly-effective teachers from the classroom,

but could be partially addressed with teachers taking on coaching responsibilities only part-time. A recent study found that pairing teachers with different strengths and weaknesses and encouraging them to coach each other is a promising strategy closely related to the coaching programs included in this analysis (Papay et al., 2016). Another approach taken by many districts has been to fold coaching into the observation component of new teacher evaluation systems. However, both theory (Herman & Baker, 2009) and case-study analyses (Kraft & Gilmour, 2016) suggest that having the same person serve as both coach and evaluator can undercut the trusting relationships needed between coaches and teachers and may result in superficial and infrequent feedback. Simply adding coaching to administrators' existing responsibilities with little training or support is unlikely to result in high-quality or sustained coaching.

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

The need for teacher buy-in presents a second major challenge for scaling-up coaching

programs. No matter the expertise or enthusiasm of a coach, coaching is unlikely to impact instructional practice if the teachers themselves are not invested in the coaching process. The programs included in this review likely benefit from the non-random sample of teachers and schools that volunteered to participate in most studies. The largest study in our sample points to the challenges of taking coaching to scale and potentially making participation mandatory. Lockwood et al. (2010) evaluate a statewide program in Florida where over 2,300 reading coaches worked with teachers across content areas to enhance literacy instruction. Across the four years they studied, effects on reading achievement were statistically significant in only two, and effects on math achievement were statistically significant in only one. Across all years, average effect sizes were extremely small, between 0.01 SD and 0.03 SD. It is not possible to determine whether these results are due to the mandatory nature of the program, the challenge of staffing such a large corps of coaches, or other factors. However, this study points to the challenges of building effective coaching programs at scale for all teachers, including some of whom may not be interested in actively participating in coaching.

The literature on schools as organizations provides some insights about how best to address the likely challenges of gaining teacher buy-in. Coaching requires teachers to be willing to open themselves to critique and recognize personal weaknesses. This openness on the part of teachers is facilitated both by a school culture committed to continuous improvement and by strong relational trust among administrators and staff members (Bryk & Schneider, 2002; Kraft & Papay, 2014). Teachers that perceive the observation and feedback cycles associated with coaching as a process intended to document shortcomings towards efforts to exit teachers may be unwilling to acknowledge a coach's critiques or take risks by experimenting with new instructional techniques (Herman & Baker, 2009; Kraft & Gilmore, 2016). This suggests that

building environments where providing and receiving constructive feedback is a regular part of teachers' professional work may be a key condition for the success of scale-up efforts.

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

Directions for Future Research

This systematic review of the teacher coaching literature reveals several ways in which scholars can improve the quality of this type of research, and highlights important directions for future work. Given the methodological inclusion criteria, the studies included in this review were overwhelmingly of high overall quality. However, there were several design and analysis practices that researchers could improve on in future studies. Many of the studies we reviewed were substantially underpowered to detect plausible effect sizes on distal outcomes such as student achievement. Studies often would 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 estimating standard errors. Finally, rates of attrition differed across studies in meaningful ways, while not all researchers tested for differential attrition or subjected their

results to robustness checks for this attrition. Future reviews may consider coding studies based on these elements of research quality as well.

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

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

- The theory of action underpinning the coaching program
- The target population of teachers, including novice versus more veteran teachers
- The fidelity of implementation of the coaching model
- The length, frequency, and total amount of coaching sessions
- The length and features of other complementary PD elements of a coaching model
- Information on how teachers and schools were recruited and compare to those that did not volunteer for a study
- The number of coaches as well as any training and support they receive
- Coach background characteristics (e.g., teaching and coaching experience, subject expertise, role in school or district)

- Estimates of the per-teacher cost of delivering the coaching program
- A clear explanation of the type of PD available to teachers and schools in the control condition
- Information about the reliability of outcome measures including observation instruments, achievement tests and self-report surveys

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

Futures studies would also benefit from examining outcomes in the year after the coaching program ends. Among the 60 studies we reviewed, only five reported outcomes from a follow-up year after coaching had ended (Allen et al., 2011; Blazar & Kraft, 2015; Garet et al., 2008; Pianta et al., 2017; Teemant, 2014). These studies present very mixed evidence about the degree to which effects are enhanced, sustained, or fade out over time. Understanding the degree to which teachers continue to implement the practices they learned with the support of a coach is essential to considering the overall costs of rolling out coaching programs at scale. Admittedly, this is not always easy to do. Maintaining the internal validity of an experimental study over time can be challenging given high rates of teacher turnover, especially in large urban 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.

In addition to improving the quality of research, the teacher coaching literature would benefit from new studies that addressed several outstanding questions. Most basically, we still know very little about the presence and scope of teacher coaching programs as they currently are being implemented across the U.S., or elsewhere around the world. We encourage researchers to advocate for the inclusion of questions about coaching activities on nationally representative

datasets in the U.S. such as the National Teacher and Principal Study and American Teacher Panel. Understanding how teacher coaching impacts teacher behaviors and student outcomes outside of the U.S., including in developing contexts, is another area in need of continued exploration. Our review also points to the relative lack of causal evidence on content-based coaching programs for subjects other than reading and literacy. The effect of coaching may differ across subject areas or for teachers with different levels of experience. Ongoing innovation in coaching practices is likely to produce new models which will present fertile areas for future research. One such example is "bug-in ear" coaching where peers or coaches provide guidance to teachers in real-time via an earpiece (Ihlo, et al., 2017; Scheeler, Congdon & Stansbery, 2010; Ottley, Coogle, Rahn, & Spear, 2017).

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

Finally, we see a critical need for studies to move beyond efficacy trials to evaluate

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

Conclusion

By pooling results from across 60 causal studies of teacher coaching programs, we find large positive effects on instruction and smaller positive effects on achievement. Effects on instruction and achievement compare favorably when contrasted with the larger body of literature on teacher PD (Yoon et al., 2007), as well as most other school-based interventions (Fryer, 2016). The growing literature on teacher coaching provides a much needed evidentiary base for future directions in teacher development policy, practice, and research. Ultimately, improving the teacher workforce will require continued innovation in in-service professional development programs. Teacher coaching models can provide a flexible blueprint for these efforts, but many questions remain about whether coaching is best implemented as smaller-scale targeted programs tailored to local contexts or if they can be taken to scale in a high-quality and cost-effective way.

References

* Indicates if a reference was included in the meta-analytic sample

^ Studies with both instruction and achievement outcomes that are included in Figure 3

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

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

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

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

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

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

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

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

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

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

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H.C. (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

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

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

*Boller, K., Del Grosso, P., Blair, R., Jolly, Y., Fortson, K., Paulsell, D.… & Kovas, M.. (2010). The seeds to success modified field test: Findings from the impact and implementation

studies. *Mathematica Policy Research*. Retrieved from https://www.mathematica-mpr.com.

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

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

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

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

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

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

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

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance:The mathematics reform in California. Teachers College Record, *102*(2), 294–343.doi:10.1111/0161-4681.00057

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

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

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

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

*Cotabish, A., Dailey, D., Robinson, A., & Hughes, G. (2013). The effects of a STEM intervention on elementary students' science knowledge and skills. *School Science and Mathematics, 113*(5), 215-226.

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

Denton, C. A., & Hasbrouck, J. A. N. (2009). A description of instructional coaching and its relationship to consultation. *Journal of Educational & Psychological Consultation, 19*

(2), 150-175. doi:10.1080/10474410802463296

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

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

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

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

*Domínguez, P., Merino, J. M., Mathiesen, M. E., Soto, M. E., & Rodríguez, C. (2016). Efecto de un programa de desarrollo profesional docente sobre la calidad de la literacidad temprana.

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

^*Downer, J. T., Pianta, R. C., Burchinal, M., Field, S., Hamre, B. K., LoCasale-Crouch, J., & Scott-Little, C. (2013). Coaching and coursework focused on teacher-child interactions during language/literacy instruction: Effects on teacher outcomes and children's classroom engagement. *unpublished paper, University of Virginia*.

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

*Early, D. M., Maxwell, K. L., Ponder, B. D., & Pan, Y. (2017). Improving teacher-child interactions: A randomized controlled trial of Making the Most of Classroom Interactions and My Teaching Partner professional development models. *Early Childhood Research Quarterly*, *38*, 57-70.

*Fabiano, G. A., Reddy, L. A., & Dudek, C. M. (2017). Teacher coaching supported by formative assessment for imposing classroom practices. *School Psychology Quarterly.*

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

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

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

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

Ganimian, A. J., & Murnane, R. J. (2016). Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research*, *86*(3), 719-755.

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

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

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

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

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

Sciences, U.S. Department of Education.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives*, *25*(4), 191-209.

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

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

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

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

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

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

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

doi:10.1111/j.1750-8606.2008.00061.x

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

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

Hill, H. C., Blazar, D., & Lynch, K. (2015). Resources for Teaching. *AERA Open, 1*(4), 1-23.

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

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

*Ihlo, T., Glover, T. A., Howell Smith, M. C., Martiin, S. D., Wu, C., & McCormick, C. M. (2017). Evaluating professional development with distance coaching for early reading RTI. Working Paper.

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

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

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

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

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

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

*Johnson, S. R., Finlon, K. J., Kobak, R., & Izard, C. E. (2017). Promoting Student–Teacher Interactions: Exploring a Peer Coaching Model for Teachers in a Preschool Setting. *Early Childhood Education Journal*, *45*(4), 461-470.

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

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

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

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

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

Koedel, C., Li, J., Springer, M.G., & Tan, L. (2017). The Impact of Performance Ratings on Job Satisfaction for Public School Teachers. *American Educational Research Journal*, *54*(2), 241-278.

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

Kraft, M.A. & Blazar, D. (2017). Individualized coaching to improve teacher practice across grades and subjects: New experimental evidence. *Educational Policy, 31*(7), 1033-1068.

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

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*.

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

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

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

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

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M.

D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*. (NCSER 2013-3000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

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

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

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

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

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

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

*McCollum, J., Hemmeter, M., & Hsieh, W. (2013). Coaching teachers for emergent literacy instruction using performance based feedback. *Topics in Early Childhood Special Education, 33*(1), 28-37. doi:10.1177/0271121411431003

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

*Milburn, T. F., Hipfner-Boucher, K., Weitzman, E., Greenberg, J., Pelletier, J., & Girolametto, L. (2015). Effects of coaching on educators' and preschoolers' use of references to print and phonological awareness during a small-group craft/writing activity. *Language, speech, and hearing services in schools, 46*(2), 94-111.

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

*Morris, P., Mattera, S., Castells, N., Bangser, M., Bierman, K., & Raver, C. (2014). Impact findings from the Head Start CARES demonstration: National evaluation of three approaches to improving preschoolers' social and emotional competence. Washington, D.C.: *Office of Planning, Research, and Evaluation, Administration for Children and*

Families, U.S. Department of Health and Human Services. Retrieved from www.acf.hhs.gov/opre

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

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

*Namasivayam, A. M., Hipfner-Boucher, K., Milburn, T., Weitzman, E., Greenberg, J., Pelletier, J., & Girolametto, L. (2015). Effects of coaching on educators' vocabulary-teaching strategies during shared reading. *International journal of speech-language pathology*, *17*(4), 346-356.

*Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Education Research Journal*, *46*(2), 532-566. doi:10.3102/0002831208328088.

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

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

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

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

^*Olson, C. B., Matuchniak, T., Chung, H. Q., Stumpf, R., & Farkas, G. (2017). Reducing achievement gaps in academic writing for Latinos and English learners in Grades 7–12. *Journal of Educational Psychology*, *109*(1), 1.

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

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

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

^*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 www.cli.org/.

^*Pianta, R., Hamre, B., Downer, J., Burchinal, M., Williford, A., LoCasale-Crouch, J., ... & Scott-Little, C. (2017). Early Childhood Professional Development: Coaching and Coursework Effects on Indicators of Children's School Readiness. *Early Education and Development*, 1-20.

^*Pianta, R. C., Burchinal, M., Jamil, F. M., Sabol, T., Grimm, K., Hamre, B. K…. & Howes, C.

(2014). A cross-lag analysis of longitudinal associations between preschool teachers' instructional support identification skills and observed behavior. *Early Childhood Research Quarterly, 29*, 144-154. doi:10.1016/j.ecresq.2013.11.006.

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

*Piasta, S. B., Justice, L. M., O'Connell, A. A., Mauck, S. A., Weber-Mayrer, M., Schachter, R. E., ... & Spear, C. F. (2017). Effectiveness of large-scale, state-sponsored language and literacy professional development on early childhood educator outcomes. *Journal of Research on Educational Effectiveness, 10*(2), 354-378.

Popova, A., & Evans, D. K., Arancibia, V., (2016). Training Teachers on the Job: What Works and How to Measure it. World Bank Policy Research Working Paper No. 7834

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental*

designs for generalized causal inference. Boston, MA: Houghton, Mifflin and Company.

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

Showers, B. (1984). Peer Coaching: A Strategy for Facilitating Transfer of Training. A CEPM R&D Report.Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Showers, B. (1985). Teachers coaching teachers. *Educational leadership*, *42*(7), 43-48.

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

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

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

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2008). Teacher professional learning and development. *Educational Practices Series – 18*. International Academy of Education.

*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 www.journals.elsevier.com

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

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

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

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

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

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement.* (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and

Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from ies.ed.gov/ncee/edlabs

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

Tables

Table 1. Characteristics of Studies Included in Meta-Analysis

	Count	Proportion
Source		
Institute Report	5	0.08
Peer-reviewed Journal	51	0.85
Working Paper	4	0.07
Year of Publication		
2006	1	0.02
2008	3	0.05
2009	4	0.07
2010	8	0.13
2011	10	0.17
2012	1	0.02
2013	3	0.05
2014	7	0.12
2015	9	0.15
2016	4	0.07
2017	8	0.13
Unknown	2	0.03
Country of Study		
Unites States	55	0.92
Chile	2	0.03
Canada	3	0.05
Research Design		
Randomized Control Trials (RCTs)	56	0.93
Quasi-experiment	4	0.07
Level of Randomization for RCTs		
Teacher	29	0.52
School	25	0.45
District	2	0.04
Teacher Sample Size		
50 or less	18	0.30
51 to 100	16	0.27
101 to 150	7	0.12
151 to 300	13	0.22
300 plus	5	0.08
Not reported	1	0.02
Coaching Model Type		
Content-Specific	40	0.67
Math	2	0.03
Reading	35	0.58
Science	3	0.05
General Practices	20	0.33
School Levels Included		
Pre-K	31	0.52
Elementary	20	0.33
Middle	15	0.25

High	7	0.12
Mode of Delivery		
In Person	47	0.78
Virtual	13	0.22
Complementary Treatment Elements		
Any Complementary Treatment	54	0.90
Group Trainings	48	0.80
Instructional Content	22	0.37
Video Library	14	0.23
Coaching Dosage (# of hours of one-on-one coaching)		
10 or less	16	0.27
11 to 20	14	0.23
21 to 30	6	0.10
30 or more	8	0.13
Not reported	16	0.27
Total PD Dosage (# of hours)		
20 or less	13	0.22
21 to 40	16	0.27
41 to 60	10	0.17
60 or more	10	0.17
Not reported	11	0.18
n	60	

Notes: School levels included and complementary treatments are not mutually exclusive.

Table 2. Pooled Effect Size Estimates of the Effect of Teacher Coaching on Instruction and Achievement

	Teacher Instruction	Student Achievement			
	Classroom Observations	All Subjects	Reading	Math	Science
All Studies	0.488***	0.178***	0.163***	0.044	0.352
	(0.056)	(0.037)	(0.032)	(0.042)	(0.242)
k[n]	186[43]	113[31]	87[26]	20[5]	6[3]
Content-Specific (All)	0.512***	0.197***	0.186***	0.050	0.352
	(0.061)	(0.041)	(0.035)	(0.041)	(0.242)
k[n]	119[27]	102[26]	78[21]	18[3]	6[3]
Content-Specific (Reading)	0.513***	0.185***	0.186***		
	(0.064)	(0.036)	(0.035)	na	na
k[n]	113[25]	82[21]	78[21]		
General Practices	0.466***	0.068	0.066		
	(0.109)	(0.056)	(0.048)	na	na
k[n]	67[16]	11[5]	9[5]		

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, *k* is the number of effect sizes and *n* is the number of studies. Cells with "na" are not estimated due to too few or no data.

Table 3. Pooled Effect Size Estimates of the Effect of Teacher Coaching on Instruction and Achievement by School Level

	Teacher Instruction	Student Achievement
	Classroom Observations	All Subjects
Pre-Kindergarten	0.480***	0.112**
	(0.072)	(0.036)
k[n]	147[27]	42[10]
Elementary School	0.559***	0.220***
	(0.161)	(0.062)
k[n]	23[10]	53[14]
Middle School	0.450***	0.175**
	(0.063)	(0.062)
k[n]	24[9]	23[11]
High School	0.492***	0.300*
	(0.121)	(0.120)
k[n]	17[5]	6[4]

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

Table 4. Meta-regression Estimates of the Relationship between Coaching Program Characteristics and Effect Sizes on Teacher Instruction

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(9)
				Panel A: Teacher Instruction				
Group Training	0.313**			0.201*				
	(0.098)			(0.084)				
Instructional Content		0.206+		0.135				
		(0.107)		(0.117)				
Video Library			-0.267**	-0.162+				
			(0.098)	(0.096)				
Total PD Features					0.102			
					(0.071)			
Virtual Coaching						-0.161		
						(0.118)		
Coaching Dosage							-0.000	
							(0.004)	
Total PD Dosage								0.001
								(0.003)
Intercept	0.239**	0.413***	0.574***	0.333***	0.338*	0.528***	0.453***	0.465***
	(0.073)	(0.075)	(0.074)	(0.100)	(0.135)	(0.067)	(0.083)	(0.110)
k[n]	186[43]	186[43]	186[43]	186[43]	186[43]	186[43]	153[34]	153[34]
				Panel B: Student Achievement				
Group Training	0.117*			0.088				
	(0.053)			(0.054)				
Instructional Content		0.084		0.061				
		(0.081)		(0.083)				
Video Library			-0.068	-0.044				
			(0.075)	(0.077)				
Total PD Features					0.051			
					(0.039)			
Virtual Coaching						-0.043		
						(0.070)		
Coaching Dosage							-0.001	
							(0.001)	
Total PD Dosage								-0.001
								(0.001)
Intercept	0.086**	0.142***	0.190***	0.092*	0.108**	0.188***	0.193***	0.255***
	(0.028)	(0.033)	(0.042)	(0.036)	(0.041)	(0.045)	(0.046)	(0.073)
k[n]	113[31]	113[31]	113[31]	113[31]	113[31]	113[31]	80[22]	80[22]

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, *k* is the number of effect sizes and *n* is the number of studies. All predictors are dichotomous except Total PD Features (1, 2 or 3) and Coaching and Total PD Dosage (hours).

Table 5. Sensitivity Analyses of the Effect of Teacher Coaching using Modified Trim and Fill Method

	Effect-Size Level		Study Level	
	Teacher Instruction	Student Achievement	Teacher Instruction	Student Achievement
	Classroom Observations	All Subjects	Classroom Observations	All Subjects
	Panel A: Unadjusted Estimates			
All studies	0.488***	0.178***	0.444***	0.131***
	(0.056)	(0.037)	(0.052)	(0.026)
k[n]	186[43]	113[31]	[43]	[26]
	Panel B: Estimates with Imputed Missing Studies			
All studies	0.343***	0.142**	0.325***	0.131***
	(0.071)	(0.039)	(0.058)	(0.026)
k[n]	226[52]	135[35]	[57]	[26]

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, *k* is the number of effect sizes and *n* is the number of studies. For effect-size level imputation we cluster effect sizes within studies according to the average number of effect-sizes per study in our primary samples. No missing studies were identified for student achievement outcomes at the study level.

Table 6. Pooled Effect Size Estimates of the Effect of Teacher Coaching by Coaching Program Size

	Teacher Instruction	Student Achievement
	Classroom Observations	All Subjects
All Studies	0.488***	0.178***
	(0.056)	(0.037)
k[n]	186[43]	113[31]
Efficacy Trials of Smaller Programs	0.631***	0.281***
	(0.083)	(0.061)
k[n]	107[26]	43[15]
Effectiveness Trials of Larger Programs	0.342***	0.099***
	(0.067)	(0.030)
k[n]	79[17]	70[16]

Notes: * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. Efficacy trials of smaller programs define by *n*(Teachers) <100 where effectiveness trials of larger programs are for *n*(Teachers)≥100. For sample size, *k* is the number of effect sizes and *n* is the number of studies.

Figures

Figure 1. Theory of Action for Teaching Coaching

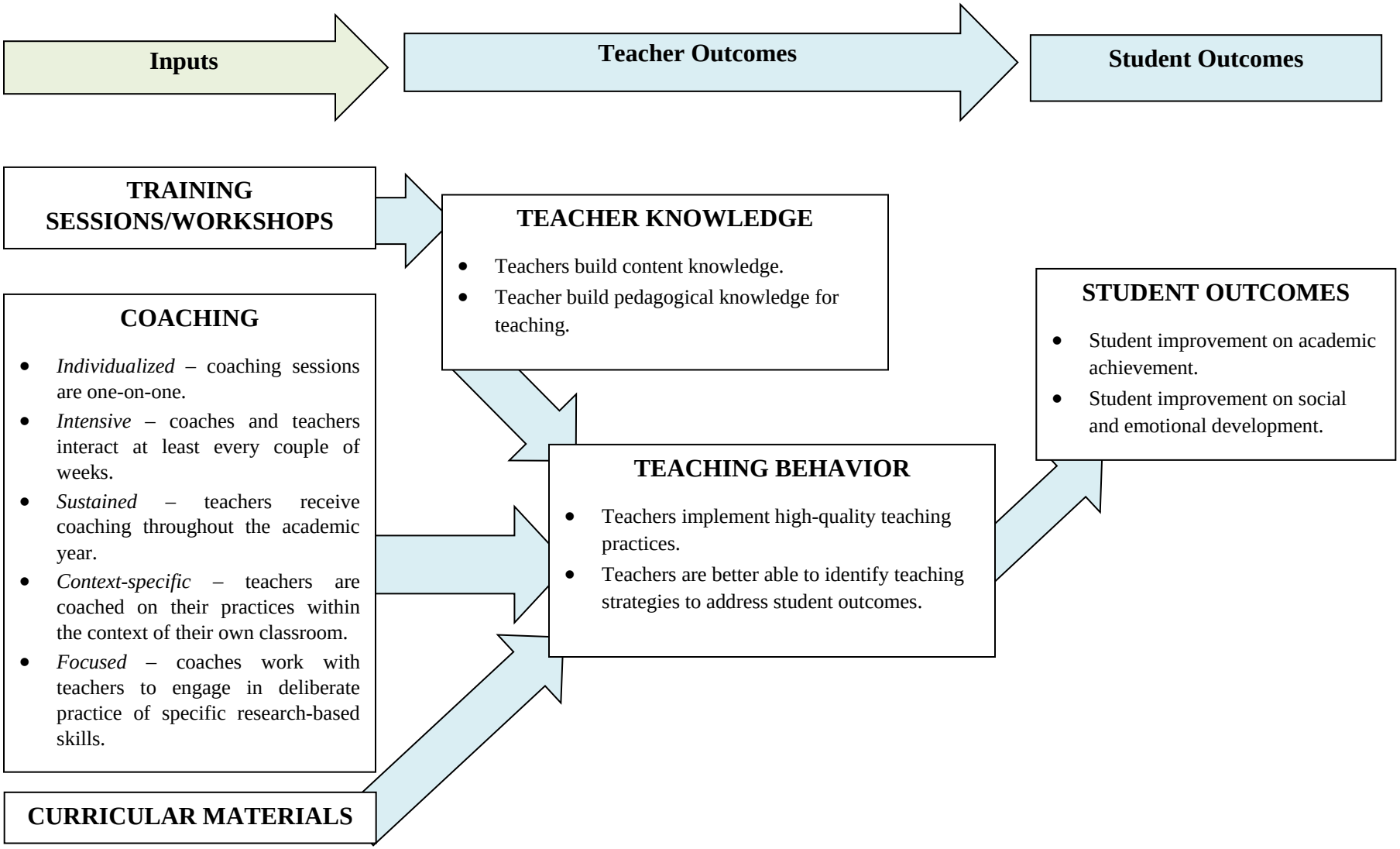


Figure 2. Kernel density plots of effect sizes for instructional and achievement outcomes.

Note. *k* = 186 for instructional outcomes and 113 for achievement outcomes

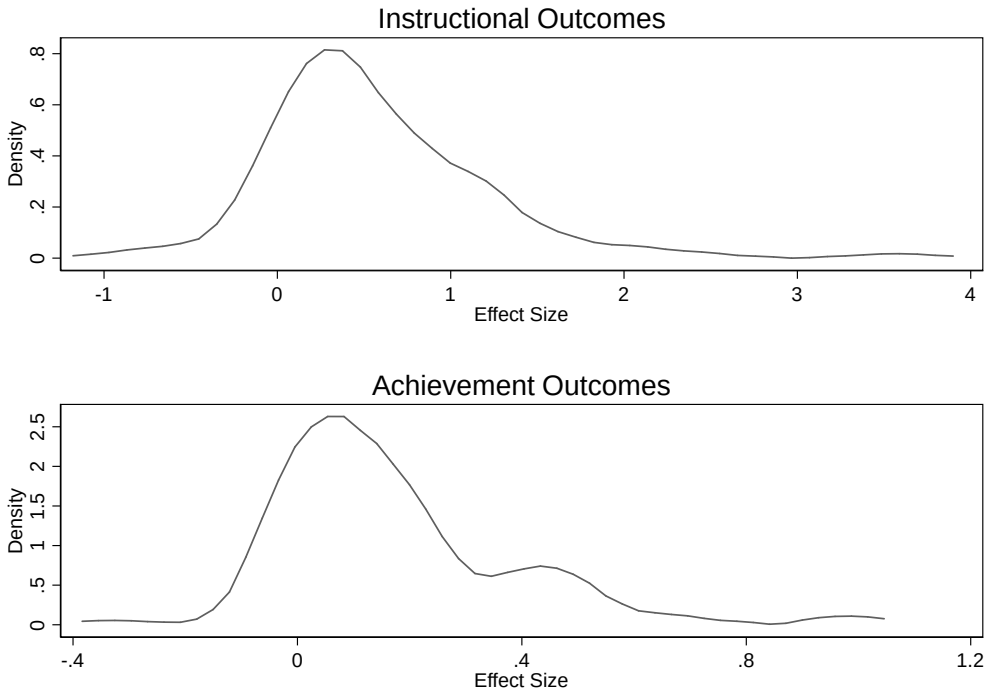


Figure 3. The relationship between coaching program effects on instruction and achievement.

Note. Data points are calculated by averaging across effect sizes for a given outcome across all effect sizes from the same research project and weighted by the average sample size. *n* = 20 studies, 16 research project.

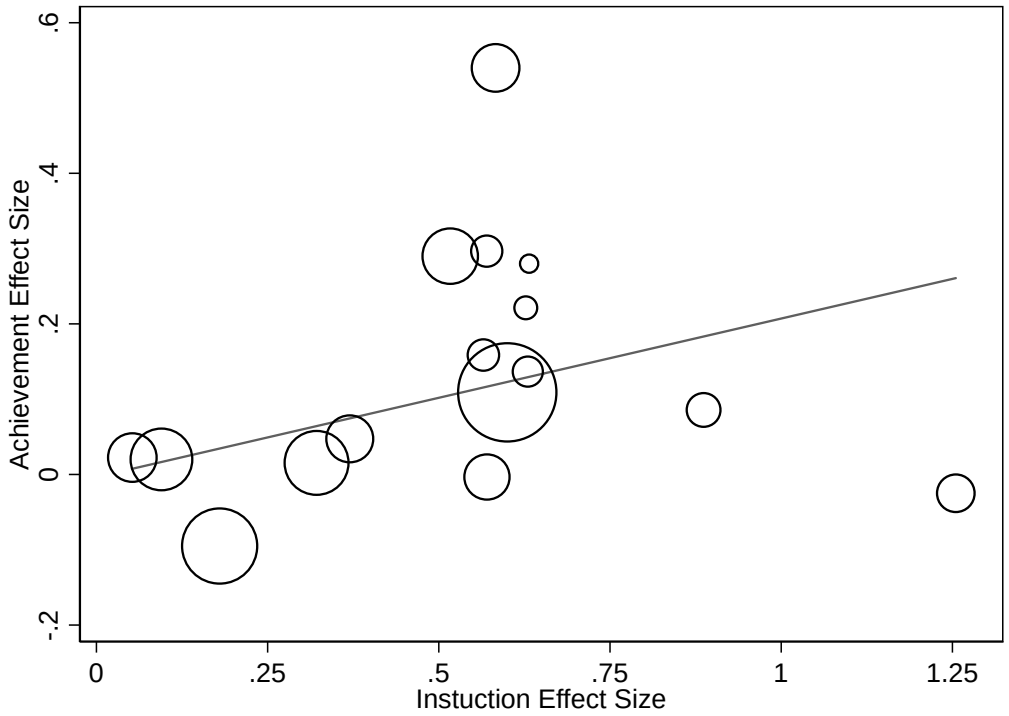
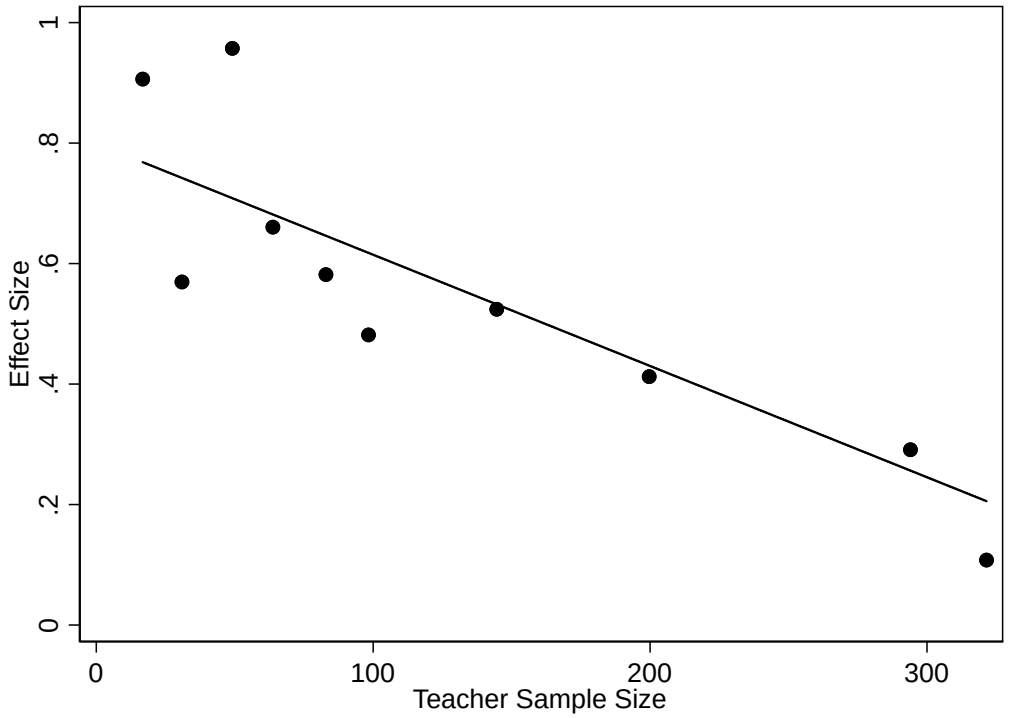


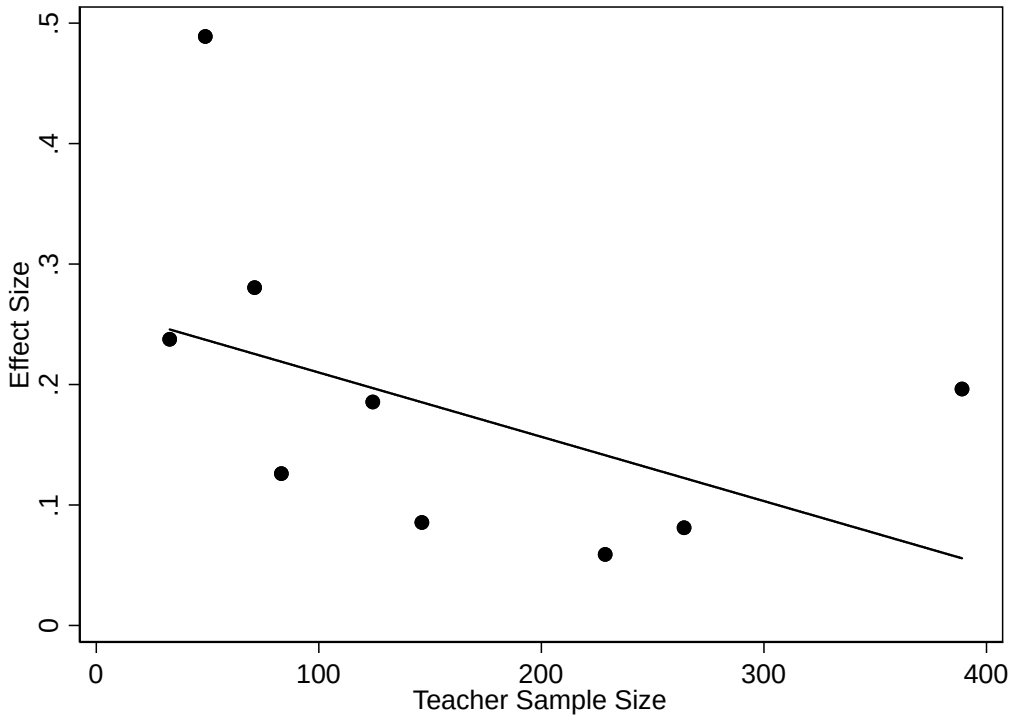
Figure 4. The relationship between effect sizes and the number of teachers participating in a study.

Note. To construct these figures, we bin test scores into deciles and plot the mean effect size within each bin. The solid line shows the best linear fit estimated on the underlying data using OLS. Panel B excludes two outliers, Campbell and Malkus et al. (2011) which reports a total teacher sample size of 1,593 and Lockwood et al. (2009) which does not report sample sizes for teachers. Panel A *n*=186, Panel B *n*= 93

Panel A: Instructional Outcomes



Panel B: Achievement Outcomes


Effect Size
0
.1
.2
.3
.4
.5
0
100
200
300
400
Teacher Sample Size

Appendix Tables

TA 1. Studies Included in Meta-Analysis

Citation	Effective Teacher Sample Size	School Level	Research Design	Outcomes	Program Type	Complementary PD Features
Abry et al. (2013)	239	Elementary	RCT	Instruction	General Instruction	Group Training, Curriculum
Allen at al. (2015)	86	Middle, High	RCT	Achievement	General Instruction	Group Training, Video Library
Allen et al. (2011)	78	Middle	RCT	Instruction & Achievement	General Instruction	Group Training, Video Library
Biancarosa, Bryk, & Dexter (2010)	259	Elementary	Diff-in-diffs	Achievement	Reading Instruction	Group Training
Bierman et al. (2008)	44	Pre-K	RCT	Achievement	Reading Instruction & General Instruction	Group Training, Curriculum
Blazar & Kraft (2015)	82	Elementary, Middle, High	RCT	Instruction	General Instruction	Group Training, Curriculum
Boller et al. (2010)	159	Pre-K	RCT	Instruction	General Instruction	Group Training
Cabell et al. (2011)	49	Pre-K	RCT		Reading Instruction	Group Training

Campell & Malkus (2011)	1593	Elementary	RCT	Achievement	Math Instruction	
Conroy et al. (2015)	53	Pre-K	RCT	Instruction	General Instruction	Group Training, Curriculum
Cotabish et al. (2013)	49	Elementary	RCT	Achievement	Science instruction	Group Training Curriculum
Dominguez et al. (2016)	48	Pre-K	RCT		Reading Instruction	Group Training
Domitrovich et al. (2009)	84	Pre-K	RCT	Instruction	General Instruction	Group Training, Curriculum
Downer et al. (na)	252	Pre-K	RCT		General Instruction	Video Library
Early et al. (2017)	311	Pre-K	RCT	Instruction	General Instruction	Video Library
Fabiano, Reddy & Dudek (2017)	89	Elementary	RCT	Instruction	General Instruction	
Fisher, Frey, & Lapp (2011)	16	Middle	RCT	Achievement	Reading Instruction	Group Training

Garet et al. (2008)	270	Elementary	RCT	Instruction & Achievement	Reading Instruction	Group Training
Garet et al. (2011)	195	Middle	RCT	Instruction & Achievement	Math Instruction	Group Training
Gregory et al. (2014)	87	Middle, High	RCT	Instruction	General Instruction	Group Training, Video Library
Hemmeter et al. (2016)	40	Pre-K	RCT	Instruction	General Instruction	Group Training, Curriculum
Ihlo et al. (2017)	389	Elementary	RCT	Achievement	Reading Instruction	Group Training Curriculum
Johnson et al. (2017)	24	Pre-K	RCT	Instruction	General Instruction	Group Training Video Library
Kraft & Blazar (2017)	50	Elementary, Middle, High	RCT	Instruction	General Instruction	Group Training, Curriculum
Landry et al. (2009)	262	Pre-K	RCT	Instruction	Reading Instruction	Group Training, Curriculum
Landry et al. (2011)	220	Pre-K	RCT	Instruction & Achievement	Reading Instruction	Group Training, Curriculum

Lockwood, McCombs, & Marsh (2010)	?	Middle	Diff-in-diffs	Achievement	Reading Instruction	
Mashburn et al. (2010)	134	Pre-K	RCT	Achievement	Reading Instruction & General Instruction	Curriculum, Video Library
Matsumara, Garnier, & Spybrook (2012)	93	Elementary	RCT	Instruction	Reading Instruction	Group Training
Matsumara, Garnier, & Spybrook (2013)	167	Elementary	RCT	Instruction & Achievement	Reading Instruction	Group Training
Matsumura et al. (2010)	73	Elementary	RCT	Achievement	Reading Instruction	
McCollum, Hemmeter, & Hsieh (2011)	13	Pre-K	RCT	Instruction	Reading Instruction	Group Training
Mikami et al. (2011)	88	Middle	RCT	Instruction	General Instruction	Group Training, Video Library
Milburn et al. (2014)	20	Pre-K	RCT	Instruction	Reading Instruction	Group Training, Curriculum
Milburn et al. (2015)	32	Pre-K	RCT	Instruction	Reading Instruction	Group Training

Morris et al. (2014)	308	Pre-K	RCT	Instruction	General Instruction	Group Training, Curriculum
Namasivayam et al. (2015)	32	Pre-K	RCT	Instruction	Reading Instruction	Group Training Video Library
Neuman & Cunningham (2009)	291	Pre-K	RCT	Instruction	Reading Instruction	Group Training
Neuman & Wright (2010)	148	Pre-K	RCT	Instruction	Reading Instruction	
Nugent et al. (2016)	124	Middle, High	RCT	Instruction & Achievement	Science instruction	Group Training, Curriculum
Olson et al. (2017)	95	Middle, High	RCT	Instruction Achievement	Reading Instruction	Group Training Curriculum
Papay et al. (2016)	136	Elementary, Middle	RCT	Achievement	General Instruction	
Parkinson et al. (2015)	130	Elementary	RCT	Instruction & Achievement	Reading Instruction	Group Training
Pianta et al. (2008)	113	Pre-K	RCT	Instruction	Reading Instruction & General Instruction	Curriculum, Video Library

Pianta et al. (2014)	252	Pre-K	RCT	Instruction	General Instruction	Video Library
Pianta et al. (2017)	252	Pre-K	RCT		General Instruction	Video Library
Piasta et al. (2017)	353	Pre-K	RCT		Reading Instruction	Group Training
Powell et al. (2010)	88	Pre-K	RCT	Instruction & Achievement	Reading Instruction	Group Training, Curriculum, Video Library
Rezzonico et al. (2015)	32	Pre-K	RCT	Instruction	Reading Instruction	Group Training Video Library
Rimm-Kaufman et al. (2014)	276	Elementary	RCT	Achievement	General Instruction	Group Training
Sailors & Price (2010)	44	Elementary, Middle	RCT	Instruction & Achievement	Reading Instruction	Group Training
Sailors & Price (2015)	120	Elementary, Middle	RCT	Achievement	Reading Instruction	Group Training
Sibley & Sewell (2011)	68	Pre-K	RCT	Instruction	Reading Instruction	Group Training, Curriculum
Teemant (2014)	36	Elementary	Diff-in-diffs	Instruction	General Instruction	Group Training

Vernon-Feagans et al. (2013)	75	Elementary	RCT	Achievement	Reading Instruction	Group Training
Vogt & Rogalla (2009)	50	Elementary, Middle, High	Diff-in-diffs	Achievement	Science instruction	Group Training
Wasik & Hindman (2011)	30	Pre-K	RCT	Instruction & Achievement	Reading Instruction	Group Training, Curriculum, Video Library
Wasik, Bond, & Hindman (2006)	16	Pre-K	RCT	Instruction & Achievement	Reading Instruction	Group Training, Curriculum
Yoshikawa et al. (2017)	76	Pre-K	RCT	Instruction	Reading Instruction & General Instruction	Group Training Curriculum
Zan & Donegan-Ritter (2014)	60	Pre-K	RCT	Instruction	Reading Instruction	Group Training

Table A2. Pooled Effect Size Estimates of the Effect of Teacher Coaching from Randomized Control Trials

	Teacher Instruction	Student Achievement	
	Classroom Observations	All Subjects	Reading
All Studies	0.451***	0.182***	0.164***
	(0.047)	(0.040)	(0.034)
k[n]	184[42]	101[28]	80[24]
Content-Specific (All)	0.512***	0.201***	0.183***
	(0.061)	(0.043)	(0.035)
k[n]	119[27]	92[24]	72[20]
Content-Specific (Reading)	0.513***	0.184***	0.183***
	(0.064)	0.035	(0.035)
k[n]	113[25]	73[20]	72[20]
General Practices	0.367***	0.052	0.055
	(0.072)	(0.070)	(0.061)
k[n]	65[15]	9[4]	8[4]

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

Table A3. Pooled Effect Size Estimates after Trimming Top and Bottom 5% of Effect Sizes

	Classroom Observations	Achievement (Pooled)	Reading Achievement
All Studies	0.453***	0.162***	0.178***
	(0.044)	(0.027)	(0.031)
k[n]	166[42]	101[29]	79[25]
Content-Specific (All)	0.526***	0.168***	0.187***
	(0.059)	(0.028)	(0.032)
k[n]	110[27]	94[26]	72[22]
Content-Specific (Reading)	0.529***	0.187***	0.187***
	(0.061)	0.032	(0.032)
k[n]	104[25]	77[22]	72[22]
General Practices	0.345***	0.11	0.11
	(0.054)	0.092	(0.092)
k[n]	56[15]	7[3]	7[3]

Notes: * p<.05, ** p<.01, *** p<.001. Trimming top and bottom 5% of effect sizes removes 16 effect size estimates for the instruction sample and 8 effect size estimates for the achievement sample. Pooled effect size estimates with robust-variance estimated standard errors reported in parentheses. For sample size, *k* is the number of effect sizes and *n* is the number of studies.