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

Providing high-quality professional development to employees is among the most important and longstanding challenges faced by organizations. Investments in on-the-job training offer large potential returns to workforce productivity. However, high-quality programs have proven difficult to develop, sustain, and scale. These challenges are particularly acute in the public education sector given the scale of the teacher labor market and the non-routine nature of the job. Every day, over 3.5 million educators face unique challenges in classrooms that often combine students with a wide range of prior knowledge, skills, and needs.

In recent years, teacher coaching has emerged as a viable alternative to the typical professional development workshops that most teachers receive. Existing handbooks and reviews of the teacher coaching literature have focused on describing the theory of action, creating typologies of different coaching models, and cataloging best implementation practices (Cornett & Knight, 2009; Devine, Meyers & Houssemann, 2013; Fletcher & Mullen, 2012; Kretlow & Bartholomew, 2010; Obara, 2010; Schachter, 2015; Stormont et al. 2015). Responding to the call by Hill, Beisiegel, and Jacob (2013), we compliment these works by conducting the first meta-analysis of studies examining the causal effect of teacher coaching on teachers’ instruction and student achievement. We focus our review narrowly on studies that employ research designs capable of supporting causal inferences. This would not have been possible only a decade ago. In 2007, a comprehensive review of the entire canon of professional development literature found only nine out of over 1,300 studies on teacher professional development were capable of supporting causal inferences (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). This finding, along with substantial investments by the Institute for Education Research (IES) and a growing movement calling for increased methodological rigor in educational research (Cook, 2001; Angrist, 2004; Murnane & Nelson, 2007; Wayne et al., 2008), served to catalyze a new wave of randomized trials evaluating coaching and other PD programs.

Our review of the literature identified 37 studies of teacher coaching programs that use both a causal research design and examine effects either on instruction (142 effect size estimates) or on student achievement (79 effect size estimates). We arrive at our pooled effect size estimates using meta-analytic methods that produce precision weighted estimates and account for the clustered nature of our data (Hedges, Tipton, & Johnson, 2010; Tanner-Smith, Tipton, & Polanin, 2016).

We find large positive effects of coaching both on teachers’ instruction and on student achievement. Pooling across studies, we find improvements on teachers’ instructional practice of 0.57 standard deviations (SD). We also estimate that coaching raised student performance on standardized tests by 0.11 SD. This estimate pools achievement tests across reading, math, and
science in order to provide a broad picture of coaching effectiveness. Disaggregating results by content area, we find similarly sized effects of reading-specific coaching programs on students’ reading achievement. Our ability to conduct this type of subgroup analysis is relatively limited for other subject areas including math and science given few studies focused on improving teachers’ math and science instruction that also measured student achievement in this same content area. A variety of sensitivity analyses suggest that these estimates cannot be explained away by missing data due to publication bias.

Coaching models differ both in their focus as well as their program features. We conduct exploratory analysis to examine whether certain program features are associated with larger or smaller mean effect sizes. We do not find any evidence of systematic differences in effect sizes based on features of the coaching model. This includes systematic differences in both instructional or achievement outcomes when coaching is combined with additional PD features, when it is delivered in person vs. remotely, as well as based on coaching dosage and total number of PD hours. However, in most cases, our standard errors are quite large and do not allow us to rule out large positive or negative effects. Uncertainty in our estimates likely is due to two reasons that reduce our statistical power. First, coaching model features vary at the study level rather than effect-size level. Second, power for meta-analysis regressions is reduced by the unbalanced distribution of these predictors in our data (Tanner-Smith, Tipton, & Polanin, 2016).

Finally, we examine a fundamental assumption underlying the theory of action for coaching and many other teacher professional development models: helping teachers improve the quality of their instructional practice will lead to improvements in student achievement (Cohen & Hill, 2000; Hanssen, 2006; Scher & O’Reilly, 2009; Weiss & Miller, 2006). Our coded meta-analysis data afford a unique opportunity to examine this critical assumption empirically using experimental studies that examine impacts on both instruction and achievement. Across our analysis we find strong supporting evidence for this hypothesis. The correlation between effect sizes on instruction and achievement across these nine data points is 0.64. Using a simple regression frame, we estimate that a one SD change in instruction associated with 0.15 SD change in achievement. This helps to illustrate why PD that results in only modest changes to teachers’ instructional practice is unlikely to lead to impacts on student achievement.

These results point to promising evidence on the effect of coaching both on teachers’ instruction and students’ academic achievement. We also finding surprising but limited exploratory evidence of lack of predictive power of specific program features, suggesting that quality of program implementation may matter just as much or more than program design. Based on these findings, we argue that future research should devote more attention to ways to bring coaching to scale.