ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-499

Instructional Coaching Personnel and Program Scalability

David Blazar
University of Maryland
College Park

Doug McNamara
University of Maryland
College Park

Genine Blue
TNTP

While teacher coaching is an attractive alternative to one-size-fits-all professional development, the need for a large number of highly skilled coaches raises potential challenges for scalability and sustainability. Collaborating with a national teacher training organization, our study uses administrative records to estimate the degree of heterogeneity in coach effectiveness at improving teachers' instructional practice, and specific characteristics of coaches that explain these differences. We find substantial variability in effectiveness across individual coaches. The magnitude of the coach-level variation (0.2 to 0.35 standard deviations) is close to the full effect of coaching programs, as identified in other research. We also find that coach-teacher race/ethnicity-matching predicts changes in teacher practice, suggesting that the relational component of coaching is key to success.

VERSION: December 2021

Suggested citation: Blazar, David, Doug McNamara, and Genine Blue. (2021). Instructional Coaching Personnel and Program Scalability. (EdWorkingPaper: 21-499). Retrieved from Annenberg Institute at Brown University: https://-doi.org/10.26300/2des-s681

Instructional Coaching Personnel and Program Scalability

David Blazar (dblazar@umd.edu)*
Doug McNamara (dmcnama1@umd.edu)
University of Maryland College Park
2311 Benjamin Building, 3942 Campus Drive, College Park, MD 20740

Genine Blue (genine.blue@tntp.org)
TNTP
500 7th Avenue, 8th Floor, New York, NY 10018

Abstract

While teacher coaching is an attractive alternative to one-size-fits-all professional development, the need for a large number of highly skilled coaches raises potential challenges for scalability and sustainability. Collaborating with a national teacher training organization, our study uses administrative records to estimate the degree of heterogeneity in coach effectiveness at improving teachers' instructional practice, and specific characteristics of coaches that explain these differences. We find substantial variability in effectiveness across individual coaches. The magnitude of the coach-level variation (0.2 to 0.35 standard deviations) is close to the full effect of coaching programs, as identified in other research. We also find that coach-teacher race/ethnicity-matching predicts changes in teacher practice, suggesting that the relational component of coaching is key to success.

Key words: teacher coaching, professional development, sustainability, race/ethnicity matching

* = corresponding author. We thank our partners and collaborators at TNTP, including Vicky Brady and Bailey Cato Czupryk, for compiling the data used in this project, and for ongoing brainstorming regarding analyses. We also thank Matthew Kraft for providing valuable feedback on the research questions and on an earlier draft of the manuscript.

Introduction and Motivation

Instructional coaching has become an attractive alternative to one-size-fits-all professional development (PD). Compared to traditional, workshop-based PD that is generally found to be ineffective (Fryer, 2017; Yoon et al., 2007), one-on-one coaching observation and feedback cycles have very large effects on teacher practice (upwards of 0.5 standard deviations [SD]) and on student test scores (upwards of 0.2 SD; Kraft et al., 2018). In fact, after reviewing experimental evidence on an array of educational interventions, Fryer (2017) found that only one-on-one, high-dosage tutoring with students had larger effects on student academic outcomes. Because tutoring is more resource intensive per student than coaching, the latter is likely a more cost-effective intervention.

Despite growing consensus on the benefits of coaching as a teacher-development tool, it is less clear how best to scale coaching programs in a way that also maintains their efficacy. Scalability and sustainability are concerns across the education research space (Slavin & Smith, 2009) but are likely to be particularly pronounced for coach-based teacher PD that relies primarily on the efficacy and skills of individual coaches. Said another way: coaches likely *are* the intervention. Blazar and Kraft (2019) provide suggestive evidence on this hypothesis by exploiting turnover of coaches across multiple cohorts of a randomized experiment. While they found large differences in the effects of the coaching program associated with individual coaches, the sample size was small (n = 5 coaches). Pooling results across all experimental studies evaluating coaching programs, Kraft et al. (2018) found effects of small-scale programs (enrolling fewer than 100 teachers and led by few coaches) that were roughly twice as large as effects of larger programs (enrolling more than 100 teachers with many more coaches). This study, too, provides speculative evidence on the importance of personnel in the coaching scale-up process, though it is an indirect test of the role that individual coaches play.

In this study, we estimate the degree of heterogeneity across individual coaches in their effectiveness at improving teachers' instructional practice, drawing on secondary data from TNTP

(formerly called The New Teacher Project). The collaboration with TNTP is appealing to examine this topic for several reasons. Because TNTP is a national, alternative-route teacher training and certification organization, our analyses leverage six years of data to examine coach effectiveness across 14 training sites (where sites generally are analogous to school districts). Thus, in addition to greatly increasing statistical power relative to prior quantitative analyses on this topic, our findings are more generalizable. Relatedly, the context and scope of TNTP's programming speaks directly to the practice and policy question at hand regarding scalability and sustainability. As described by Kraft et al. (2018), many rigorous evaluations of coaching programs have been conducted under best-case scenarios, with relatively small samples of teachers, small numbers of coaches, and where coaches often were the program designers (sometimes also members of the research team). Yet, in real-world settings, districts need to hire much larger corps of coaches and to recruit them from broad labor pools. TNTP's programming closely reflects this context.

To estimate heterogeneity in coach effectiveness, we take a value-added approach that is similar to the teacher effectiveness literature (Hanushek & Rivkin, 2010). Specifically, we predict teachers' observed quality of instruction at the end of coaching as a function of baseline observation measures (hence "value-added") and additional covariates that aim to capture the primary avenues through which coaches are matched with teachers (e.g., site, certification area). While a randomized trial—in which coaches are randomly assigned to teachers—would provide stronger evidence of heterogeneity in effectiveness across individual coaches, we find that our value-added approach generally passes falsification tests that estimate the "effect" of coaches on measures that they should not impact (i.e., background teacher demographic characteristics). We focus on coaching cycles and data collected over the summer prior to individuals' first year as full-time teacher of record. This is TNTP's pre-service training period and, thus, the time during which the organization hires a very large corps of coaches and where data are collected systematically across sites.

Overall, we find substantial variability across coaches in terms of changes in teacher practice. A 1 SD increase in coach effectiveness is associated with a 0.2 to 0.35 SD increase in multiple dimensions of teaching practice. Results of coach-level variation are similar when we nest coaches within sites to test for variation at each level, as well as when we estimate coach-level variation across each of the four largest training sites. These patterns suggest that it is the coach—and not the support, training, and oversight provided by each site—that likely matters most. Our estimates of coach-level variation in changes in teacher practice are roughly two-fifths to three-quarters of the *full* effect of coaching programs, on average (Kraft et al., 2018).

To further aid schools and districts looking to implement or scale coaching models through targeted recruitment and coach development, we also examine whether observable characteristics of coaches predict changes in teacher performance. We find positive associations for coach-teacher race/ethnicity matches. These patterns align with theoretical discussion of coaching as a relational activity (Joyce & Showers, 1981), while also suggesting that recruitment of a diverse pool of coaches and screening for coaches' interpersonal skills may be one of the best strategies for scale-up.

Theoretical Framework on Performance Heterogeneity

Longstanding lines of theoretical and empirical work point to substantial heterogeneity in the efficacy of personnel and labor pools. In this paper, our work builds most directly from research showing variation in the effectiveness of individual teachers, with studies consistently showing effects on student test scores of roughly 0.2 student-level SD (Hanushek & Rivkin, 2010) and even larger effects on social-emotional outcomes including student engagement in class activities (0.3 SD; Blazar & Kraft, 2017). Our work also aligns with newer lines of research that find substantively meaningful variation across principals (Grissom et al., 2015) and guidance counselors (Mulhern, 2019) in terms of effects on student outcomes.

The education sector is an appealing context for examining variation in effectiveness across personnel, as there exist clear and measurable indicators of productivity: namely, student outcomes (Todd & Wolpin, 2003), as well as observable measures of the quality of teachers' classroom instruction that predict changes in student outcomes (Bell et al., 2012; Kane et al., 2011). Measuring personnel productivity vis-à-vis performance outcomes also has longstanding discussion in the health sector, with doctors linked to patient outcomes (Safran et al., 1998), and in the economics and management literature on firms (Holmstrom & Milgrom, 1991).

In light of this prior work, we expect that coaches, too, will exhibit variability in performance when linked to key teacher outcomes. After all, at their core, coaching programs are meant to be individualized, driven by the needs of individual teachers and one-on-one development work implemented by individual coaches. In their pioneering work describing the theory of action underlying coaching models, Joyce and Showers (1981) note that coaching "represents a continuing problem-solving endeavor between the teacher and the coach..." that relies on "...a collegial approach to the analysis of teaching for the purpose of integrating mastered skills and strategies into: (a) a curriculum, (b) a set of instructional goals, (c) a time span, and (d) a personal teaching style" (p. 170). Aligned to this perspective, additional researchers and practitioners describe coaching as a relational endeavor driven primarily by coaches' "people skills," including building relationships and trust with teachers, and differentiating support for individual teachers' needs (Denton & Hasbrouck, 2009; Wong & Nicotera, 2006).

In turn, we argue that it is not the mere existence of variation in effectiveness across individual coaches that matters, but rather the *magnitude* of that variation. For coaching to be a viable intervention across states, districts, and schools, it is necessary to identify, recruit, hire, train, support, and retain large corps of coaches, potentially pulling current, highly effective teachers out of classrooms to serve in these roles (Darling-Hammond, 2017). Like teachers, coaches also need to be trained and supported

(Stoetzel & Shedrow, 2020), requiring additional resources. Thus, substantial variation in performance across coaches could undermine growing interest in coaching as a primary—if not *the* primary—PD tool. In the 2007-08 school year, roughly 57% of public schools nationally had at least one coach (National Center for Education Statistics [NCES], 2008), compared to 66% of schools in 2015-16 (NCES, 2016). Yet, to our knowledge, only Blazar and Kraft (2015, 2019) quantitatively examine mean differences in effects of individual coaches when linked to teacher outcomes. Because their analyses focus on a small sample of five coaches, they are unable to estimate the true underlying variance in coach effects. In other words, the sample of five coaches may not reflect coach-level variability in larger district, state, or national populations where coaching programs are implemented.

Two personnel and performance management questions related to scaling coaching programs also are relevant to the current topic: What are the key domains of coach characteristics that explain their effects on teacher performance? How can these skills be leveraged for recruitment and screening of, and professional learning for coaches? Here, there is a small but growing literature base. By and large, coaches tend to be expert teachers with a demonstrated track record of success in the classroom, who often enter the role through a career ladder; coaches may come from within a school or district, or from another context (Darling-Hammond, 2017; Wenner & Campbell, 2017). In terms of the specific characteristics and skills of potential coaches to look for, Connor (2017) hypothesizes three areas of effectiveness. First, there must be a strong interpersonal relationship between the coach and teacher. Coaches and teachers who communicate and collaborate more effectively may experience bigger rewards from the coaching relationship. Second, a coach's knowledge of effective teaching and coaching practices may affect teaching outcomes. Similarly, more effective coaches may have content-specific knowledge which they use in the coaching relationship. Knowledge of effective teaching practices plays a direct role in ensuring high-quality observation-feedback cycles. Third, the types of

tools (e.g., modeling, providing direct feedback, video observation, etc.) and technologies (e.g. online vs. in-person coaching, bug-in-ear real-time coaching, etc.) a coach uses may matter.

Empirically, scholars have started to operationalize domains of coach skill in survey instruments and observation tools to capture the quality of coach-teacher interactions (e.g., Howley et al., 2014), examine variability in how coaches instantiate these practices in their work with teachers (e.g., Shannon et al., 2021), and link coach characteristics and practices to teacher outcomes (e.g., Marsh et al., 2012; Yopp et al., 2019). For example, in the context of a math coaching program in Tennessee, Russell et al. (2020) found that a 1 SD change in the depth and specificity of coaches' conversations with teachers was associated with a 0.2 SD increase in the quality of teachers' instruction. However, much of this work has been conducted in small samples, generally with no more than 30 coaches. Further, because this literature base is quite new, many of the theorized domains of coach effectiveness have not been linked to changes in teacher practice, particularly in samples that can lead to generalizable conclusions. While we are not able to examine all hypothesized domains of coach effectiveness, we are able to provide suggestive evidence on some of the key skills highlighted in the theoretical literature; and, we examine heterogeneity in coach effectiveness across a number of U.S. states and school districts.

Research Design

In this study, we ask: (1) *To what extent do individual coaches vary in their contributions to changes in teachers' instructional practice?* (2) *To what extent do observable characteristics of coaches (i.e., years of coaching experience, demographic matches with teachers) explain their effects on teacher practice?*

Empirical Strategy

To answer these questions, we draw on the teacher effectiveness and value-added literatures (Hanushek & Rivkin, 2010) to specify a production function of the following form:

$$OBSERVATION_{ijst}=\beta_0+\beta_1 OBSERVATION_{ijsc(t-1)}+\beta_2 I_{j(t-1)}+\delta_{st}+(\mu_{js}+\varepsilon_{ijst}) \quad (1)$$

where the outcome of interest is the end-of-coaching observation score for teacher *i* working with coach *j* in site *s* and year *t*. The key feature of our model is that we control for a baseline measure of the outcome, $OBSERVATION_{ijs(t-1)}$, captured at the beginning of the training period and prior to the start of coaching. Controlling for a baseline measure allows us to estimate changes in teacher practice associated with individual coaches and, most importantly, to account for bias due to non-random sorting of coaches to teachers. To this same end, we further control for baseline teacher characteristics (i.e., gender, race/ethnicity) and certification area, included in the vector, $I_{j(t-1)}$, as well as site-year fixed effects, δ_{st}. According to TNTP, these are the primary avenues and characteristics that drive coach-teacher matches.

Our primary estimate of interest comes from the coach or coach-year random effect—μ_{js} in equation (1) or μ_{jst} in alternative specifications—which provides a model-based estimate of the variation in changes in teacher practices associated with individual coaches. Coach-level random effects consider coach effects as stable across years, while coach-year random effects allow for variation across years. As shown below, we find that both sets of estimates are quite similar. Our random-effects, multilevel model shrinks the coach or coach-year effects back towards the mean based on the precision of those estimates, driven primarily by the number of teachers with whom an individual coach works (mean = 8.2 teachers per coach/year, SD = 2.5). In some models, we nest the coach-year random effect within a site-year random effect—moving δ_{st} from the fixed to the residual portion of the model—in order to examine whether coaches versus the sites within which they work are a primary driver of changes in teacher outcomes.

Data and Sample

We fit our models using data collected by TNTP across six years (2014 through 2019) and 14 summer training sites. Our primary sample includes a census of pre-service teachers (n = 3,526) and coaches (n = 317) with whom TNTP worked during this time period. In Table 1, we show that this

sample of teachers is roughly two-thirds female, one-quarter Black, and two-fifths White. (Twenty percent of teachers did not report race/ethnicity information.) These characteristics are more diverse than national characteristics of teachers (NCES, 2020), but are aligned with characteristics of teachers who go through alternative-certification programs—including TNTP—that often operate in urban settings and often have a goal of decreasing barriers to entry into the profession for historically marginalized groups (NCES, 2016; Shen, 1997). Demographic characteristics of coaches are similar to those of teachers: roughly two-thirds are female, one-quarter are Black, and half are White; three-quarters have one year of experience coaching for TNTP.

Trained evaluators rated teachers' instructional practice multiple times over the course of the summer using TNTP's observation rubric (TNTP, 2017). This rubric includes three dimensions of practice, each of which is scored on a scale from 1 (Ineffective) to 3 (Developing): (i) *Culture of Learning* asks whether all students are engaged in the work of the lesson from start to finish, and focuses on the extent to which teachers maximize instructional time and maintain high expectations for student behavior; (ii) *Essential Content* asks whether all students are engaged in content aligned to the appropriate standards of their subject and grade, and focuses on the extent to which teachers plan and deliver content accurately and clearly; and (iii) *Demonstration of Learning* asks whether all students demonstrate that they are learning, and focuses on the extent to which teachers check for student understanding and respond to student misunderstandings. Observers, hired by TNTP, participated in rater training during which they rated no fewer than seven full-length instructional videos followed by three to four "check in" points to rate and discuss additional lesson videos or co-observe in classrooms. Overall, observers receive about 40 to 50 hours a year of observation practice. We standardized observation scores to have a mean of 0 and SD of 1.

All three domains of teaching practice have been linked to student test score growth in other TNTP-led research projects (TNTP, 2018). Our own analyses, shown in Table 2, also provide

evidence that these scores capture the underlying construct of interest. Lesson-level intraclass correlations (ICC) range from 0.36 to 0.49, and are similar to other studies in which trained observers score the quality of teachers' instruction (Bell et al., 2012; Hill et al., 2012). Our analyses focus on these lesson-level scores as the outcome of interest, though we also note that adjusted teacher-level ICCs—that accumulate information across lessons—are substantially higher, ranging from 0.55 to 0.69. Measurement error in our dependent variables can limit the precision of our estimates, but will not lead to attenuation bias, as is the case with measurement error in independent variables.

In most instances, observations are conducted and scored by the teachers' coach. While this setup closely matches the purpose of coaching models—organized around observation and feedback cycles led by the coach—it could bias our estimates of variation in coach effectiveness given that the coach is both the key input and the one responsible for measuring outcomes. At the same time, we find that, amongst a set of sites and years in which lessons were observed both by the teachers' own coach and another observer, interrater agreement rates are comparable to other studies in which trained observers score the quality of teachers' instruction (Bell et al., 2012; Hill et al., 2012): 70% for *Culture of Learning*, 66% for *Essential Content*, and 51% for *Demonstration of Learning* (see Table 2). Further, in a set of robustness tests that focus only on lessons observed by outside raters, we find that variation in coach effectiveness is larger than in the full sample.

Findings

Heterogeneity in Effectiveness Across Coaches

We begin, in Table 3, by showing the variation in coach effectiveness as measured by changes in each of the four measures of teaching practice (the three individual dimensions and the composite measure), pooling across all sites and years. We find that a 1 SD increase in coach effectiveness is associated with a roughly 0.2 SD increase over the course of the summer in the composite measure of teacher practice (0.19 SD for the coach random effect, and 0.22 SD for the coach-year random

effect). Our estimates of variability in coach effectiveness are similar for *Culture of Learning* and for *Essential Content*, and slightly larger for *Demonstration of Learning* (0.24 to 0.29 SD).

In Appendix Table 1, we re-estimate coach effects using a subset of site-years in which a rater other than teachers' own coach observed and scored their instruction. We find that the variation in coach effectiveness often is larger than in the full sample: roughly 0.3 SD for the composite measure of practice, roughly 0.22 SD for *Culture of Learning* and *Essential Content*, and 0.33 to 0.35 SD for *Demonstration of Learning*. The latter dimension of practice is where inter-rater agreement rates between a teacher's own coach and another rater were lowest (see Table 2). Therefore, it appears that we are underestimating variation in coach effectiveness by using scores rated by teachers' own coach. That said, as we proceed with our results, we rely on the largest possible sample in order to maximize precision and generalizability. Here and in Table 3, estimates of coach and coach-year variation are quite similar, though the latter often are estimated more precisely. Therefore, we focus primarily on coach-year random effects in the rest of our analyses.

In Table 4, we present additional estimates that examine the extent to which variation in coach effectiveness is driven by specific sites. Even though all sites operate under a common TNTP coaching model and management structure, each site hires its own coaches and provides training, support, and management to them. Given this, one might expect to see variation in changes in teacher practices and coach effectiveness across sites. However, overall, we find that it is the coach and not the site that appears to be primarily responsible for changes in teacher practice. In column 1, we nest coach-years within site-years in our random effects structure, finding negligible and non-significant variation at the site-year level (0.02 SD) and similar variation at the coach-year level (0.2 SD). In the next four columns, we disaggregate coach effects for the four largest training sites, each of which has a sample of at least 30 coaches when pooling across available years of data. Estimates of the coach-year variation range from 0.17 SD to 0.23 SD.

Coach Characteristics that Predict Changes in Teacher Practice

Knowing that coaches vary substantially in their effects on teacher practice begs the question: What characteristics, knowledge, and skills of coaches explain these differences? TNTP's administrative records include background data on coaches that align with theory on some key dimensions of coach quality: (i) *years of coaching experience with TNTP* serves as a proxy for the accumulated knowledge and skills coaches build in their work over time, while (ii) *teacher-coach demographic matches* may increase the strength of interpersonal relationship between coaches and teachers (Connor, 2017).

In Table 5, we examine whether these characteristics predict changes in teacher outcomes, adding these characteristics to the fixed portion of our model outlined in equation (1) above. Here, we expand our analyses to focus on all four measures of teaching practice, given robust theoretical discussion about how race/ethnicity-matching can be particularly beneficial for building culturally relevant and responsive classroom environments (Ladson-Billings, 1995). We estimate relationships using equation (1), with observable coach characteristics added to the fixed portion of the model. Given the composition of our teacher and coach samples (see Tables 1) that are comprised primarily of Black and White individuals, we focus on three race/ethnicity categories: Black, White, and non-Black/non-White. We exclude teachers and coaches who are missing information on race/ethnicity or gender. In the top panel, we start with models that include main effects of individual coach characteristics; in the bottom panel, we interact coach demographic characteristics with teacher demographic characteristics to examine the role of matching.

In both the top and bottom panels, we do not find evidence that increased experience as a TNTP coach is associated with larger changes in teacher practice. Estimates linking a dummy indicator for having a coach in their third year of experience or higher (compared to having a first- or second-year coach) for the composite measure of instructional practice and *Culture of Learning* both are positive

but not statistically significantly different from zero. The TNTP coach sample is primarily composed of early-career coaches, and so we may be underpowered to detect effects. Nonetheless, the point estimates are small.

We find some evidence that having a male coach is related to changes in *Culture of Learning* (top panel), though the estimate (0.09 SD) only is statistically significant at the $p = 0.1$ threshold. In turn, we also examine male teacher-coach matches (bottom panel), finding positive point estimates when predicting all four teaching practice measures; however, none of these estimates is statistically significantly different from zero. We observe similar patterns for the main effect of having a Black coach: all four point estimates are positive but none are statistically significantly different from zero.

Comparatively, we find that assignment of a Black teacher to a Black coach is associated with a 0.18 SD increase in the composite measure of effective instruction, and a 0.22 SD increase in *Culture of Learning*. These estimates compare Black teachers with a Black coach to their Black peers with a White coach, White teachers with a non-White coach, and non-Black/non-White teachers with a White coach or a non-Black/non-White coach. Results are almost identical when we change the reference category. We also control for the main effect of having a Black coach; though not shown in Table 5, none of these estimates are statistically significantly different from zero (consistent with the patterns from the top panel). We also find that Black teachers assigned to a non-Black/non-White coach outperform their peers (0.26 SD for the composite measure and 0.25 SD for *Culture of Learning*). We do not find any statistically significant relationships of race-matching for White coaches working with White teachers.

Identification Check

The internal validity of our findings relies on the assumption that teacher-coach assignments are random, conditional on covariates included in the model (i.e., baseline measure of teaching practice, teacher demographics, and site-year and certification area fixed effects). We assess this

assumption in Appendix Table 2 by conducting a falsification test that estimates the "impact" of coaches on observable background teacher characteristics (i.e., gender, race/ethnicity), still controlling for a baseline measure of the outcome, and site-year and certification area fixed effects. Positive and statistically significant coach "effects" here do not invalidate our value-added methodology, but rather point to potential sorting bias that is not fully accounted for with the set of available covariates (Goldhaber & Chaplin, 2015). We find that the coach-level variation is zero or very close to zero when predicting each of the race/ethnicity dummy variables.[1] When predicting teacher gender, we observe non-zero variation at the coach or coach-year level, but the estimate is roughly a third as large as when predicting teacher practices. These patterns suggest that our covariates likely have accounted for potential sorting bias, of coaches to teachers generally (relevant for analyses of individual coaches effects) and of coaches to teachers of different races or ethnicities (relevant for analyses of race/ethnicity matches).

Discussion and Conclusion

Using a value-added approach similar to the teacher effectiveness literature, we present evidence that individual coaches are the key ingredient for success of coaching programs. Across a range of models and specifications, we observe substantial variation across coaches in how teachers improve their instructional practice. The magnitude of coach-level variation as measured by changes in teacher practice is particularly large when compared to the full effect of coaching programs. We find that a 1 SD increase in coach effectiveness is associated with a 0.2 to 0.35 SD increase in multiple dimensions of teaching practice, whereas meta-analytic estimates indicate that coaching programs, on

[1] Random effects models have known challenges when estimates are close to zero (Harville, 1977). For example, when the estimated variance approaches zero, the standard error is undefined (i.e., estimates in Appendix Table 2 predicting dummy indicators for Black teacher and White teacher). To confirm that our estimates are true zeros, we estimated results to 10 decimal places, finding similar results.

average, improve teacher practice by roughly 0.5 SD (Kraft et al., 2018). In other words, variation in effectiveness across individual coaches explains almost the full effect of coaching programs.

Further, aligned to the work of other scholars (Connor, 2017; Denton & Hasbrouck, 2009; Joyce & Showers, 1981; Wong & Nicotera, 2006), we theorize that there are multiple potential mechanisms that might explain differences in coach effectiveness: the knowledge and skill that coaches bring to their work with teachers, coaches' interpersonal relationships with a given teacher, and the types of tools the coaches use. While use of administrative records means that we have a limited set of variables to capture these varied skills, we find initial evidence that the second avenue related to interpersonal relationships may be key to coach effectiveness and coaching program success. We find that Black teachers assigned to a Black or to a non-Black/non-White coach outperformed their peers in terms of changes in instructional practice; these differences are driven primarily by changes in classroom climate and cultural components of high-quality teaching. Drawing from the theoretical literature on teacher-student racial matches (Ladson-Billings, 1995), we argue that these patterns may be driven by the unique interpersonal relationships that teachers and coaches can develop when they have similar shared experiences and understandings. Comparatively, additional years of coaching experience—a proxy for the background knowledge and skill that coaches bring to their work—is not associated with increased teaching quality.

To confirm and extend these findings, future research might estimate coach effects under experimental conditions, where coaches are randomly assigned to teachers. This design then could be paired with more extensive data collection on the various theorized dimensions of coach quality and skill, with each dimension then linked to teacher outcomes. Identifying coach practices and skills that improve teachers' delivery of rigorous content and teachers' work with students around that content would help build on our findings. Our estimates of coach-level variation are largest when predicting *Demonstration of Learning*, which focuses on these teacher practices; however, we did not find that

observable coach characteristics available in our data predicted changes in this measure. Future research might also link individual coaches and their skills to student-level outcomes, in addition to teacher-level ones. Estimates of coach effects on student outcomes almost certainly will be smaller than coach effects on teacher-level outcomes, given that the former are more distal than the latter in the instructional improvement process. That said, the magnitude of variability in coach effectiveness associated with changes in teaching practices (upwards of 0.35 SD) suggests that relationships may further translate into changes in student outcomes.

Ultimately our findings have broader implications for schools and districts interested in expanding their coaching programs. Currently, school districts spend approximately $18 billion on PD each year (Education Next, 2018) for the 3.5 million full-time teachers in the United States (NCES, 2020). However, these dollars generally are found to have very little, if any, return on investment (Fryer, 2017; Yoon et al., 2007). Coaching provides an attractive alternative, achieving some of the largest impacts on teacher and student outcomes across all of the education intervention literature (Fryer, 2017; Kraft et al., 2018). Further, the overall costs of coaching programs are comparable to other PD offerings. Knight and Skrtic (2021) find that the primary ingredients of coaching programs are the coach salary and teacher time, with average costs ranging from $5,300-$10,500 per teacher, per year (adjusted to 2021 dollars). The literature on costs of more traditional teacher PD is older, but suggests that expenditures are quite similar, at $3,100 to $11,700 per teacher per year (also adjusted to 2021 dollars; Miles et al., 2004). In other words, coaching is likely to be substantially more cost effective than traditional PD. Further, because coaching purposefully is individualized and differentiated, it likely makes sense to provide coaching only to some teachers who need it most and only in some school years. This approach would further decrease the overall coaching program costs from the district perspective.

At the same time, adopting and scaling instructional coaching in lieu of traditional PD is a risky proposition without knowing how to identify effective coaches—whose salary is the key cost driver of coaching programs (Knight & Skrtic, 2021)—and how to recruit, train, and support more of them. Based on findings from our study, we offer several recommendations for policy and practice. First, our value-added methodology offers one way to identify effective coaches. Like in the teacher effectiveness realm, these measures could be used to make ongoing personnel decisions related to retention and salary. Second, positive relationships between coach-teacher demographic matches and changes in teaching practice suggest that recruitment efforts may focus on building a diverse corps of coaches whose characteristics match demographics of local teacher workforces. We recognize that efforts to diversify coach workforces may work against simultaneous efforts to diversify the teacher workforce, given that coaches often are current or former teachers in the same or a nearby district (Darling-Hammond, 2017; Wenner & Campbell, 2017). That said, large effects of virtual coaching programs (e.g., Allen et al., 2011) suggest that hiring could occur outside of a local area. Further, we hypothesize that mechanisms underlying coach-teacher demographic matches likely are related to interpersonal relationships. Thus, school districts—and researchers—may focus on designing instruments to screen and train this skill set, particularly in instances where matching coach and teacher demographics may not be possible.

Rigorous empirical evidence indicates that coaching should be at the forefront of instructional improvement efforts. Scaling and sustaining these programs is doable (Kraft et al., 2018), but will require strategic planning that focuses primarily on building a corps of highly skilled coaches.

References

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*(6045), 1034-1037.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

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

Blazar, D., and Kraft, M. A. (2019). Balancing Rigor, Replication, and Relevance: A Case for Multiple-Cohort, Longitudinal Experiments. *AERA Open*, *5*(3).

Connor, C. M. (2017). Commentary on the special issue on instructional coaching models: Common elements of effective coaching models. *Theory into Practice*, *56*(1), 78-83.

Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice?. *European Journal of Teacher Education*, *40*(3), 291-309.

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

Education Next. (2018, June 12). EdStat: $18 Billion a Year is Spent on Professional Development for U.S. Teachers. *Education Next*. Retrieved from: http://www.educationnext.org/edstat-18-billion-year-spent-professional-development-u-s-teachers/

Fryer, J., Roland G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments* (Vol. 2, pp. 95-322). North-Holland.

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein Falsification Test": Does it really show teacher value-added models are biased?. *Journal of Research on Educational Effectiveness*, *8*(1),

8-34.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis*, *37*(1), 3-28.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267-71.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320-338.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., ... & Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, *17*(2-3), 88-106.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*, *7*, 24.

Howley, A. A., Dudek, M. H., Rittenberg, R., & Larson, W. (2014). The development of a valid and reliable instrument for measuring instructional coaching skills. *Professional Development in Education*, *40*(5), 779-801.

Joyce, B. R., & Showers, B. (1981). Transfer of training: The contribution of "coaching". *Journal of Education*, *163*(2), 163-172.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587-613.

Knight, D. S., & Skrtic, T. M. (2021). Cost-effectiveness of instructional coaching: Implementing a design-based, continuous improvement model to advance teacher professional development. *Journal of School Leadership*, *31*(4), 318-342.

Kraft, M. A., Blazar, D., and Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A Meta-Analysis of the Causal Evidence: *Review of Educational Research, 88*(4) 547-

588.

Ladson-Billings, G. (1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory into practice*, *34*(3), 159-165.

Marsh, J. A., McCombs, J. S., & Martorell, F. (2012). Reading coach quality: Findings from Florida middle schools. *Literacy Research and Instruction*, *51*(1), 1-26.

Mulhern, C. (2019). Beyond teachers: Estimating individual guidance counselors' effects on educational attainment. *Cambridge, MA: Harvard University. Retrieved January*, *26*, 2020.

National Center for Education Statistics. (2008). *School and Staffing Survey*. Retrieved from: https://nces.ed.gov/pubs2009/2009321/tables/sass0708_2009321_s12n_06.asp

National Center for Education Statistics. (2016). *National Teacher and Principal Survey*. Retrieved from: https://nces.ed.gov/surveys/ntps/tables/Table_5_042617_fl_school.asp

National Center for Education Statistics. (2020). *Characteristics of Public School Teachers*. https://nces.ed.gov/programs/coe/indicator_clr.asp

Miles, K. H., Odden, A., Fermanich, M., & Archibald, S. (2004). Inside the black box of school district spending on professional development: Lessons from five urban districts. *Journal of Education Finance*, *30*(1), 1-26.

Russell, J. L., Correnti, R., Stein, M. K., Thomas, A., Bill, V., & Speranzo, L. (2020). Mathematics coaching for conceptual understanding: Promising evidence regarding the Tennessee math coaching model. *Educational Evaluation and Policy Analysis, 42*(3), 439–466.

Safran, D. G., Taira, D. A., Rogers, W. H., Kosinski, M., Ware, J. E., & Tarlov, A. R. (1998). Linking primary care performance to outcomes of care. *Journal of Family Practice*, *47*, 213-220.

Shannon, D. K., Snyder, P. A., Hemmeter, M. L., & McLean, M. (2021). Exploring Coach–Teacher Interactions Within a Practice-Based Coaching Partnership. *Topics in Early Childhood Special Education*, *40*(4), 229-240.

Shen, J. (1997). Has the alternative certification policy materialized its promise? A comparison between traditionally and alternatively certified teachers in public schools. *Educational Evaluation and Policy Analysis*, *19*(3), 276-283

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500-506.

Stoetzel, L., & Shedrow, S. (2020). Coaching our coaches: How online learning can address the gap in preparing K-12 instructional coaches. *Teaching and Teacher Education*, *88*.

TNTP. (2017). *TNTP Core Teaching Rubric*. Retrieved from: https://tntp.org/assets/documents/TNTP_Core_Teaching_Rubric_2017-18.pdf

TNTP. (2018). *The Opportunity Myth: Technical Appendix*. Retrieved from: https://files.eric.ed.gov/fulltext/ED590222.pdf

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3-F33.

Wenner, J. A., & Campbell, T. (2017). The theoretical and empirical basis of teacher leadership: A review of the literature. *Review of Educational Research*, *87*(1), 134-171.

Wong, K., & Nicotera, A. (2006). Peer coaching as a strategy to build instructional capacity in low performing schools. In K. Wong and S. Rutledge (Eds.), *System-wide efforts to improve student achievement*. Greenwich, CT: Information Age Publishing.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033. *Regional Educational Laboratory Southwest (NJ1)*.

Yopp, D. A., Burroughs, E. A., Sutton, J. T., & Greenwood, M. C. (2019). Variations in coaching knowledge and practice that explain elementary and middle school mathematics teacher change. *Journal of Mathematics Teacher Education*, *22*(1), 5-36.

Tables

Table 1. Characteristics of Teachers and Coaches

	Teachers	Coaches
Demographics		
Female	0.66	0.67
Male	0.30	0.21
Missing Gender	0.03	0.12
Asian	0.03	0.03
Black	0.26	0.25
Hispanic	0.04	0.04
White	0.40	0.52
Multiple Races/Ethnicities	0.06	0.04
Missing Race/Ethnicity	0.20	0.12
Certification Area		
Early Childhood Education	0.07	NA
Elementary School	0.24	NA
English Language Arts (ELA)	0.11	NA
Math	0.08	NA
Science	0.09	NA
Social Studies	0.01	NA
English as a Second Language	0.04	NA
Special Education	0.15	NA
Foreign Language	0.01	NA
Missing Certification Area	0.20	NA
Coaching Experience with TNTP		
Total yrs.	NA	1.36
1 yr. Experience	NA	0.74
2 yrs. Experience	NA	0.19
3 or more yrs. Experience	NA	0.07
Persons (*n*)	3,526	317

Table 2. Descriptive Statistics for Observation Scores

Observation Scores (1 to 3 Scale)	Univariate Statistics				Reliability		
	Last Score		First Score		Lesson-Level ICC	Teacher-Level Adjusted ICC	Inter-Rater Agreement
	Mean	SD	Mean	SD			
Composite	2.51	0.50	0.25	0.53	0.49	0.69	NA
Culture of Learning	2.51	0.63	2.28	0.68	0.47	0.68	70%
Essential Content	2.72	0.52	2.50	0.63	0.31	0.55	66%
Demonstration of Learning	2.31	0.70	1.97	0.71	0.36	0.61	51%

Note: ICC = intraclass correlation. Following a generalizability framework, teacher-level ICCs are adjusted for the median number of lessons per teacher. Inter-rater agreement is not calculated for the composite, as researchers (not observers) calculated the composite as an average of the other three dimensions of teaching practice.

Table 3. Standard Deviation of Coach-Level Variation, Pooling Across Sites

	Composite	Culture of Learning	Essential Content	Demonstration of Learning
Coach-Year Random Effect	0.219***	0.217***	0.202***	0.288***
	(0.023)	(0.024)	(0.026)	(0.023)
Coach Random Effect	0.191***	0.198***	0.170***	0.241***
	(0.024)	(0.025)	(0.028)	(0.025)
Teachers (*n*)	3,526	3,526	3,526	3,526
Coach-Years (*n*)	430	430	430	430
Coaches (*n*)	317	317	317	317

Notes: Each estimate comes from a separate multilevel model of teachers' end-of-summer observation score on baseline scores for all three dimensions of practice, teacher gender and race/ethnicity, certification area fixed effects, and site-year fixed effects. *** z >3.29, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to the reader, but they should be interpreted with caution.

Table 4. Standard Deviation of Coach-Level Variation on Composite Measure of Instructional Practice, by Site

	All Sites	Site 1	Site 2	Site 3	Site 4
Site-Year Random Effect	0.023	NA	NA	NA	NA
	(0.085)				
Coach-Year Random Effect	0.196***	0.219***	0.218***	0.171~	0.225***
	(0.024)	(0.042)	(0.050)	(0.089)	(0.074)
Teachers (*n*)	3,526	873	719	326	399
Coach-Years (*n*)	430	96	90	45	46
Coaches (*n*)	317	59	47	36	32

Notes: Estimates in each column come from separate multilevel models of teachers' end-of-summer observation score on baseline scores for all three dimensions of practice, teacher gender and race/ethnicity, certification area fixed effects, and site-year or year fixed effects. ~ z >1.64, *** z >3.29, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to the reader, but they should be interpreted with caution.

Table 5. Predictive Power of Coach Characteristics

	Composite	Culture of Learning	Essential Content	Demonstration of Learning
		Main Effects		
3 or more yrs. Experience	0.046	0.073	-0.010	-0.010
	(0.104)	(0.108)	(0.112)	(0.116)
Black Coach	0.081	0.082	0.065	0.047
	(0.052)	(0.054)	(0.056)	(0.059)
Non-Black/Non-White Coach	0.050	0.062	0.038	-0.010
	(0.066)	(0.069)	(0.071)	(0.074)
Male Coach	0.061	0.091~	-0.02	0.015
	(0.050)	(0.052)	(0.054)	(0.057)
		Demographic Matching		
3 or more yrs. Experience	0.036	0.064	-0.017	-0.012
	(0.104)	(0.108)	(0.112)	(0.116)
Black Teacher*Black Coach	0.181~	0.222*	0.073	-0.016
	(0.107)	(0.111)	(0.119)	(0.116)
Black Teacher*Non-Black/Non-White Coach	0.261*	0.246~	0.168	0.088
	(0.131)	(0.136)	(0.146)	(0.141)
White Teacher*White Coach	-0.135	-0.158	-0.086	0.007
	(0.097)	(0.101)	(0.108)	(0.104)
Male Teacher*Male Coach	0.058	0.024	0.083	0.026
	(0.079)	(0.082)	(0.088)	(0.085)
Coaches (*n*)	265	265	265	265
Teachers (*n*)	2,591	2,591	2,591	2,591

Notes: Estimates in each panel and column come from separate multilevel models that include coach-year random effects. All models control for baseline scores for all three dimensions of practice, teacher gender and race/ethnicity, certification area fixed effects, and site-year fixed effects. In models with teacher-coach demographic match indicators, main effects of coach and teacher demographics also included as controls. *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1.

Appendix Tables

Appendix Table 1. Standard Deviation of Coach-Level Variation in Sample where Raters are not Teachers' Coach

	Composite	Culture of Learning	Essential Content	Demonstration of Learning
Coach-Year Random Effect	0.301***	0.224***	0.211***	0.352***
	(0.051)	(0.060)	(0.058)	(0.052)
Coach Random Effect	0.288***	0.220***	0.222***	0.327***
	(0.056)	(0.059)	(0.058)	(0.059)
Teachers (*n*)	749	749	749	749
Coach-Years (*n*)	92	92	92	92
Coaches (*n*)	81	81	81	81

Notes: Each estimate comes from a separate multilevel model of teachers' end-of-summer observation score on baseline scores for all three dimensions of practice, teacher gender and race/ethnicity, certification area fixed effects, and site-year fixed effects. *** z >3.29, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to the reader, but they should be interpreted with caution.

Appendix Table 2. Falsification Tests

	Female	Asian	Black	Hispanic	White
Coach-Year Random Effect	0.074***	0.020*	0.000	0.014	0.000
	(0.015)	(0.008)	--	(0.012)	--
Coach Random Effect	0.068***	0.011	0.000	0.020*	0.000
	(0.014)	(0.013)	--	(0.008)	--
Coaches (*n*)	317	317	317	317	317
Teachers (*n*)	3,526	3,526	3,526	3,526	3,526

Notes: Each estimate comes from a separate multilevel model of teachers' end-of-summer observation score on baseline scores for all three dimensions of practice, certification area fixed effects, and site-year fixed effects. When female is the outcome, a missing gender dummy and race/ethnicity dummies also are included as controls; when race/ethnicity dummies are the outcomes, a missing race/ethnicity dummy and gender dummies are included as controls. "–" indicates that the relevant parameter could not be estimated. * $z > 1.96$, *** $z > 3.29$, where z equals the ratio of a given random effects parameter estimate to its standard error. These z-scores do not correspond precisely to p-values as in a traditional linear regression framework, as the confidence interval for a random effect estimate is not symmetric around the estimate (random effect estimates have a lower bound of zero). These traditional markers of significance are included as an aid to the reader, but they should be interpreted with caution.