COLLEGE OF EDUCATION
TRANSFORMING EDUCATION FOR GOOD
RESEARCH REPORT
May 2022

Instructional Coaching Personnel and Program Scalability

David Blazar,[1] Doug McNamara,[1] & Genine Blue[2]
[1]University of Maryland, [2]TNTP

Instructional coaching is an attractive alternative to one-size-fits-all teacher training and development in part because it is purposefully differentiated: programming is aligned to individual teachers' needs and implemented by an individual coach. But, how much of the benefit of coaching as an instructional improvement model depends on the specific coach with whom a teacher works? Collaborating with a national teacher training and development organization, TNTP, we find substantial variability in effectiveness across coaches in terms of changes in teachers' classroom practice (0.43 standard deviations). The magnitude of coach effectiveness heterogeneity is close to average coaching program effects identified in other research. These findings suggest that identifying, recruiting, and supporting highly skilled coaches will be key to scaling instructional coaching programs.

Introduction

Instructional coaching is an attractive alternative to one-size-fits-all teacher training and professional development. Compared to traditional, workshop-based programs that generally are 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]) that translate into meaningful impacts for students (upwards of 0.2 SD on test scores; 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 academic outcomes. Because tutoring is more resource intensive per student than coaching, the latter likely is a more cost-effective intervention. Instructional coaching also has gained substantial popularity across the U.S., with the number of coaches per student roughly doubling between 2000 and 2010 (Domina et al., 2015) and continued growth of programs since then. In the 2015-16 school year, 66% of public schools nationally had at least one coach (National Center for Education Statistics [NCES], 2016), compared to 57% of schools in 2007-08 (NCES, 2008).

Despite growing interest in and consensus on the benefits of instructional coaching as a teacher training and development tool, it is less clear how best to scale programs in a way that also maintains their efficacy. Scalability is a concern across the education research space (Slavin & Smith, 2009) but is likely to be particularly pronounced for coach-based teacher training and development that relies primarily on the efficacy of individual coaches. While coaching programs differ to some extent in design features, the intervention model is defined by coaches' engagement with teachers in one-on-one instructional improvement processes that include time-intensive classroom observation and feedback cycles (Joyce & Showers, 1981). The success of these efforts in improving the quality of teachers' classroom instruction is, thus, thought to depend on the knowledge, skills, and interpersonal relationship-building that individual coaches bring to their work (Connor, 2017; Denton &

Hasbrouck, 2009; Wong & Nicotera, 2006). For instructional coaching to be a viable intervention across teacher training organizations, districts, and schools, it is necessary to identify, recruit, and hire very large corps of highly skilled coaches, potentially pulling current, highly effective teachers out of classrooms to serve in these roles (Darling-Hammond, 2017). As such, substantial variability in performance across individual coach*es* could undermine efforts to make coach*ing* a primary—if not the primary—teacher training and development tool.

In this paper, we estimate the degree of performance heterogeneity across individual coaches in their effectiveness at improving the quality of teachers' instructional practice, drawing on secondary data from TNTP (formerly called The New Teacher Project). Variation in effectiveness across coaches can be thought of as heterogeneity in impacts within an instructional coaching program. The collaboration with TNTP is appealing to examine this topic for several reasons. Because TNTP is a national, alternative-route teacher training, certification, and development organization, our analyses leverage six years of data to examine heterogeneity in effectiveness across 317 coaches, working in 14 training sites (generally equivalent to school districts), and spread across 13 states. Thus, in addition to greatly increasing statistical power relative to the few prior quantitative analyses on this topic with five coaches (Blazar & Kraft, 2015, 2019), our findings are more generalizable. Relatedly, TNTP's programming speaks directly to the practice and policy question at hand regarding scalability. As described by Kraft et al. (2018), many evaluations of coaching programs have been conducted under best-case scenarios, with small numbers of coaches who often were the program designers. Yet, in real-world settings, teacher training organizations, districts, and schools need to recruit and hire much larger corps of coaches from broad labor pools. TNTP's programming closely reflects this context, where the organization hires up to 130 total coaches per year and up to 20 coaches per site and year.

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 when data are collected

systematically across sites. During pre-service training, TNTP trainees split their time between teaching in summer-school classrooms and largely coach-based learning activities that we describe in more detail below (Menzes & Maier, 2014). While the literature base on instructional coaching to date has focused on in-service training and development (Kraft et al., 2018), there is growing attention to coaching in pre-service training in traditional certification programs (e.g., Britton & Anderson, 2010; Cohen et al., 2020) and alternative-route certification models like TNTP's (e.g., Foote et al., 2011; Kaufman et al., 2020). Pre-service instructional coaching also aligns with literature on the role of cooperating or mentor teachers during field placements (e.g., Matsko et al., 2020; Ronfeldt et al., 2018; Ronfedlt et al., 2020). We still refer to pre-service trainees as teachers, given that the structure of the alternative-route certification program means that they are actively teaching students enrolled in summer school. Further, our analyses examine measures of instructional practice during lessons taught by these individuals.

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 examine changes in teachers' observed quality of instruction from the beginning to the end of coaching associated with individual coaches, controlling for covariates that aim to capture the primary avenues through which teachers are matched with coaches (e.g., site, certification area). While a randomized trial—in which teachers are randomly assigned to coaches—would provide stronger evidence of heterogeneity in effectiveness across coaches, we find that our value-added model passes placebo and falsification tests that estimate the "effect" of coaches (i) after randomly assigning teachers to coaches that they did not actually work with, and (ii) on immutable teacher characteristics that they cannot impact. In our primary analyses, we limit our sample to sites and years where there was a programmatic decision to hire outside raters to score the quality of teachers' instruction. Although the nature of observation and feedback cycles generally means that coaches themselves collect teacher performance

measures, using coach-collected data could bias our estimates of interest since coaches generate the outcomes and they are the inputs. In supplemental analyses, we find that patterns generalize to the full sample to sites and years, no matter who rated teachers' instruction.

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 0.43 SD increase in multiple dimensions of instructional quality measured on TNTP's observation instrument, including the extent to which teachers provide content aligned to appropriate standards for grade and subject, and teachers' supports for students to engage meaningfully in classroom activities. A 2 SD increase in coach effectiveness—or the difference between having a coach at the 84^{th} versus the 16^{th} percentile in the performance distribution—is associated with a 0.86 SD increase in teachers' observed quality of instruction. We further show that the magnitude of the coach-level variation is similar across the largest TNTP sites and when we parse out the unique contribution of coaches versus sites. This suggests that coach effectiveness heterogeneity is not simply a product of where they work and the oversight of each site.

Policymakers and practitioners who are considering building or scaling instructional coaching programs likely are most interested in coach effects on student outcomes. After all, students are the ultimate beneficiary of instructional improvement programs. Our dataset does not include student outcomes due to the summer school context where state or district tests generally are not administered. That said, differences in teachers' instructional quality for having a highly effective versus a less effective coach that we document are quite large relative to average effects of instructional coaching programs on measures of teachers' classroom practice documented in other research (roughly 0.5 SD); these average effects on teaching quality further translate into large student-test score gains (roughly 0.2 SD; Kraft et al., 2018). We therefore infer that teachers' students are likely to feel the impact of highly effective versus less effective coaches as well, and that identifying, recruiting, and supporting highly skilled coaches will be key to the scalability of instructional coaching programs.

Framework and Motivating Literature on Performance Heterogeneity

We hypothesize that individual coaches likely vary to some degree in their effectiveness at improving desired educational outcomes, namely the quality of teachers' instruction. Whether the magnitude of that variation is large versus small has important implications for scalability.

We come to this hypothesis based on the nature of instructional coaching as an individualized intervention—which we briefly describe above and return to below—as well as from broader lines of theoretical and empirical work that point to substantial heterogeneity in the efficacy of personnel and labor pools. The most immediate link is to the teacher effectiveness literature, where studies consistently show that teachers differ not only in the quality of their classroom instruction (Bell et al., 2012; Hill et al., 2015; Kane & Staiger, 2012), but also in their subsequent impacts on students' test scores and social-emotional development (Kraft, 2019; Hanushek & Rivkin, 2010). Our analyses also align with newer lines of research that find substantively meaningful variation across principals (Grissom et al., 2015) and guidance counselors (Mulhern, 2019) in their effects on student outcomes. Outside of the education sector, examining personnel productivity vis-à-vis performance outcomes 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).

One appealing framework derived from this literature is that the effectiveness of individual personnel can be estimated by way of their impacts on key beneficiaries—such as teachers, counselors, and principals linked to student outcomes, and, in this paper, coaches linked to teacher outcomes. A second learning is that we must consider not just whether individuals differ in their performance, but more importantly the magnitude of that variation. In studies where teachers have been randomly assigned to students, teacher effect estimates on student test scores range from 0.15 to 0.25 SD; teacher effect estimates on dimensions of students' social-emotional development often are larger (Blazar, 2018; Kane & Staiger, 2008; Kraft, 2019; Nye et al., 2004). This means that, on average,

assignment to a teacher at the 84th percentile of effectiveness moves the medium-performing student to roughly the 60th percentile, relative to students' peers assigned to a teacher at the 50th percentile in the performance distribution. These differences are quite large as benchmarked against students' average yearly test-score gains, the effect of varied educational interventions, and policy-relevant gaps in achievement between students from different backgrounds (Hill et al., 2008). Findings related to performance heterogeneity across teachers have led to general consensus that teachers are by far the most important within-school resource that we can provide to students.

Applying a framework of performance heterogeneity to instructional coaches, it is possible that there may be similar—if not greater—variability in performance as has been observed for other labor pools such as teachers. After all, at their core, coaching programs are meant to be individualized, driven both by the needs of individual teachers with whom they work and one-on-one development work implemented by individual coaches. In their pioneering work describing the theory of action underlying instructional 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, others 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).

Exploratory analyses of coach characteristics and practices indicates that, indeed, the instructional coaching experience can differ for teachers depending on the coach with whom they work. Across different coaching programs and models, teachers identify differences in their rapport with coaches, and coaches themselves report variation in the specific activities they engage in with teachers (e.g., reviewing assessment data, reflecting with teachers on their instruction, goal and action

planning; Marsh et al., 2012; Russell et al., 2020; Shannon et al., 2021; Yopp et al., 2019). Some of these coach characteristics and activities link to teacher outcomes, including their content knowledge and observed quality of instruction.

To our knowledge, only Blazar and Kraft (2015, 2019) quantitatively examine variation in effectiveness of individual coaches at improving teacher outcomes in a similar fashion as the teacher effectiveness literature. Here, we differentiate between who coaches are and the things that coaches do with teachers (descried in the previous paragraph) from the impacts they have on desired outcomes. Certain coach characteristics and coaching activities likely explain variation in coach effectiveness—a topic that we return to in our discussion. In their study, Blazar and Kraft found substantial differences in average treatment effects of the coaching program across multiple cohorts of their randomized experiment, with large positive effects in the first cohort but null effects in two subsequent cohorts. Exploratory analyses suggested that differential treatment effects across cohorts likely were attributable in part to turnover of coaches and differences in coach effectiveness. On average, the teachers of the most effective coach scored roughly 1.2 SD higher than the teachers of the least effective coach on instructional quality measures derived from classroom observations, as well as 0.7 SD higher on student-reported measures of classroom experiences. At the same time, the small sample of five coaches cannot speak to an underlying population distribution of coach effectiveness. It may be that large differences in effectiveness across five coaches are due to sampling idiosyncrasies and potential outliers. Thus, a primary goal of our analyses is to examine heterogeneity in coach effectiveness at improving teachers' instructional practice in a much larger sample.

A related line of research, often drawing on large samples in administrative datasets, considers links between cooperating or mentor teachers in pre-service field placement settings and mentee teacher outcomes. By hosting pre-service teachers in their classrooms, cooperating or mentor teachers can take on coaching-like work, including modeling instructional practice, observing mentees when

they take over lessons for periods of time, and providing feedback on instruction (Matsko et al., 2020). Findings from these studies identify benefits to teacher outcomes of having a cooperating or mentor teacher who is more instructionally effective (Bastian et al., 2020; Goldhaber et al., 2020; Ronfeldt, Brockman, & Campbell, 2018), with some suggestion that these benefits are driven by coaching activities in addition to the other roles the cooperating or mentor teacher serves (e.g., job-search support, general encouragement; Matsko et al., 2020; Ronfeldt et al., 2018; Ronfedlt et al., 2020).

However, a key distinction between cooperating or mentor teachers versus instructional coaches is the programmatic structure. While cooperating or mentor teachers have a general goal of improving a teacher's practice, these roles are described in the literature as lacking a core definition (Matsko et al., 2020). In contrast, instructional coaching is built on a robust theoretical literature base on the instructional improvement process that is guided by core observation and feedback cycles, even if details of those cycles are adapted by individual coaches and for individual teachers (Joyce & Showers, 1981). Because instructional coaching—like cooperating or mentor teacher placements—is personnel focused, we hypothesize *some* degree of variation in effectiveness across coaches. But how much? Is the coaching model robust to who implements it? Or, is it that coaches *are* the intervention?

The TNTP Coaching Model and Pre-Service Training Context

We explore variation in coach effectiveness in the context of TNTP's instructional coaching model for pre-service teachers. TNTP is an alternative-route teacher certification entity, which has trained and certified over 50,000 teachers since opening its doors in 1997. Like other alternative-route teacher certification programs, TNTP partners with school systems to recruit prospective teachers largely from local labor pools, with the goal of filling local teacher vacancies in hard-to-staff subject areas and schools (Walsh & Jacobs, 2007). The nature of the alternative-route certification program means that training is condensed into five to seven weeks prior to becoming a full-time teacher of record, and the practicum component occurs in summer school classrooms. We describe the

alternative-route, pre-service training as context for where coaching activities take place, but note that this paper does not aim to evaluate or compare alternative versus traditional certification pathways.

Aligned to longstanding calls for and trends in teacher education and training reform—within which alternative certification programs have played a key role (Wilson, 2014)—in 2012 TNTP shifted its programming to focus more intentionally on a targeted set of foundational teaching skills, and on providing time for teachers to practice and receive directed feedback on their implementation of these skills in real-world classrooms (Menzes & Maier, 2014). Our study focuses on this post-2012 time period. The prioritized set of instructional skills include: clear delivery of lessons, maintaining high academic and behavioral expectations, and maximizing instructional time. These elements of instructional practice—and the quality of teachers' implementation of them—are instantiated in a classroom observation instrument developed at TNTP that guides formative assessment and feedback, as well as summative evaluations to determine whether or not prospective teacher candidates earn provisional licensure and certification. In our study, we use this instrument to capture the quality of instructional practice outcome measures (see discussion below).

Attention to practice and feedback as key resources for developing teaching skill align closely with the theory of action underlying instructional coaching programs (Joyce & Showers, 1981). On average over the course of TNTP's summer training period, teachers spend at least 32 hours working with an instructional coach. (Teachers also have field placements in the classroom of a cooperating or mentor teacher, though this person does not simultaneously fill the role of a TNTP instructional coach.) Training starts with coaches showing teachers examples of what effective classroom environments look like, both through videotapes of exemplar lessons and modeling. Then, coaching observation and feedback cycles begin with three core components: active observations, direct and specific feedback, and immediate practice. Coaches typically engage in the process through visits to teachers' (and their cooperating or mentor teachers') classes where they observe instruction. Coaches

may also explicitly model a particular teaching skill or guide teachers in more subtle ways, including in-the-moment feedback (e.g., holding up signs or whispering to the teacher). Following a classroom visit, coaches meet with teachers for debriefing sessions to provide "bite-sized" feedback on one or two observed elements of instruction. These feedback points stem from the classroom observation and are meant to help teachers improve in their very next lesson. A goal for the feedback process is to provide teachers with concrete and manageable steps that they can address that day or the next day. Teachers may practice this new technique in front of their coach during the debrief session. For additional details on TNTP's pre-service training and coaching model, see Menzes and Maier (2014).

While TNTP coaching and pre-service training operates under a common organizational model, individual coaches are the program implementers and they do so with guidance from site managers. In many instances, sites are large school districts; in other instances, sites are state agencies that partner with TNTP to recruit and train prospective teachers for placement in different local education agencies across the state. Each summer, central office staff for each site hire coaches, pulling both from pools of TNTP-trained teachers and local educators. Coaches are expected to have a minimum of two years of successful teaching experience in high-need subject areas, familiarity with the instructional standards associated with the school district in which they are serving, and demonstrated ability to support teacher trainees in developing the teaching techniques emphasized in TNTP's training model. In the spring and early summer, coaches receive up to 40 hours of training from site leads who often were coaches themselves in prior years. Coach training led by sites generally includes an overview of the coaching model, practicing coaching, and observing and scoring the quality of classroom instruction. Following training, coaches work individually with teachers, providing guidance and support aligned to their observations of teachers' instruction in summer-school classrooms and their perceptions of teachers' most immediate needs.

Research Design

In this study, we ask: *To what extent do individual coaches vary in their effectiveness at improving the quality of teachers' instructional practice?*

Data and Sample

To answer this question, we rely on data collected by TNTP across six years (2014 through 2019) and 14 summer training sites located across 13 states. A key feature of the data is that we can directly link coaches to teachers and, then, to teacher outcomes. Within each site, teachers and coaches work across several schools that serve as field placements. However, our dataset does not include links between teachers and school or cooperating teacher placements.

The data include information on 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-route teacher certification programs that often operate in urban settings with 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. Coaches may have coaching experience outside of TNTP, which we are not able to capture in the administrative records available for this study.

Our outcome measures of interest are dimensions of the quality of teachers' classroom practice. Trained evaluators observed and rated teachers' instruction multiple times over the course of the summer using TNTP's rubric (TNTP, 2014). Before scoring teachers' instruction, observers participated in 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 received about 40 to 50 hours a year of observation practice. The rubric includes three dimensions of instructional practice, each of which is scored on a quality 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. (For additional details, see TNTP [2013, 2014].) We also created a composite measure of the quality of teachers' practice that is an average of these three dimensions. We standardized scores to have a mean of 0 and SD of 1.

All of these dimensions of teaching practice have been linked to student test score growth in other TNTP-led research projects (TNTP, 2018) and in an external validation study (McEachin et al., 2018). In Table 2, we also provide evidence that these scores capture the underlying construct of interest as defined by the instrument—i.e., the quality of classroom instruction—as opposed to construct-irrelevant sources of variation such as raters. Lesson-level intraclass correlations (ICC) range from 0.36 to 0.49, and are similar to other studies in which trained observers scored teachers' instruction (Bell et al., 2012; Hill et al., 2012). While our analyses focus on changes in these lesson-level scores, we also note that adjusted teacher-level ICCs that accumulate information across lessons are higher, ranging from 0.55 to 0.69. For a subset of lessons scored by more than one rater, inter-rater agreement rates are comparable to other studies (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).

In many instances, a teacher's coach conducted and scored observations, which aligns with the setup of coaching models that are organized around observation and feedback cycles led by the coach. At the same time, use of these scores 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. Therefore, in our primary analyses, we focus on a subset of five sites and 11 site-year combinations where there was a programmatic decision to have outside raters score the quality of teachers' instruction on the final observation of the summer. One reason that some sites hired external raters to observe and score this summative observation is because that score most directly links to provisional licensure decisions before teachers entered the classroom as a full-time teacher in the fall.[1] As we describe below, the summative, post-coaching observation score is the metric that we use to define and measure coach effectiveness heterogeneity.

In Table 1, we refer to this group as the "rater-not-coach" sample. This subsample looks similar to full sample with regard to teacher certification areas, gender, and most of the race/ethnicity groups, and teacher certification, as well las coaches' years of experience with TNTP. Differences in other background characteristics of teachers and coaches are due to differences in which sites contribute to the "rater-not-coach" sample.

[1] Our strategy for estimating coach effectiveness also relies on a baseline observation score captured before the start of coaching activities. We do not make any restrictions on who scored this baseline observation for two reasons. First, only one site made a systematic decision to have external raters—rather than coaches—score teachers' instruction for both the pre- and post-coaching periods. More importantly, because the baseline measure was captured prior to the start of coaching activities, having a coach rate that lesson should not introduce bias into our estimates of coach effectiveness. We provide suggestive empirical evidence in support of this claim using data from the one site where external raters scored both the pre- and post-coaching observations, and where we also have scores from teachers' coach (n = 100 teachers and 12 coaches). For this sample, we find very similar estimates of coach effectiveness heterogeneity when we use pre- and post-coaching teacher observation scores from external raters (0.32 SD) versus when we use a pre-coaching score from the coach and a post-coaching score from an external rater (0.29 SD). These estimates are similar to other site-specific estimates of coach effectiveness heterogeneity that we present in Appendix Table 2. Predictably though, we are underpowered to detect whether or not these estimates are statistically significantly different from zero nor from each other when limiting the sample to just one site and to the subset of teachers within this site that were simultaneously observed both by their coach and an external rater.

Empirical Strategy

Guided by the teacher effectiveness literature, we estimate variability in effectiveness across coaches in terms of improvements in the quality of teachers' instruction by specifying a production function of the following form:

$$\Delta OBSERVATION_{ijst} = \beta I_{j(t-1)} + \delta_{st} + (\mu_j + \varepsilon_{ijst}) \quad (1)$$

In this model, the outcome of interest is the change in classroom observation score from the beginning to the end of coaching for teacher *i* working with coach *j* in site *s* and year *t*. Focusing on a change score accounts for teachers' starting point at the beginning of the summer and aims to minimize selection 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 that are 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 teacher-coach matches.

An alternative value-added approach common in the teacher effectiveness literature is to control for a baseline measure of the outcome on the right-hand side of the equation. This setup allows researchers to flexibly model the relationship between baseline and outcome measures. In contrast, examining change scores as the outcome of interest assumes that the relationship between baseline and outcome measures is linear and that the correlation between them is 1. At the same time, including a baseline measure on the right-hand side of the equation can lead to attenuation bias if that variable is measured with error. In a teacher value-added model focused on student test-score growth, it is reasonable to assume that measurement error is small. Student assessments generally are constructed to have reliability at or above 0.9. This is not the case, though, when focusing on lesson-level instructional quality scores, where we—like other scholars—document intraclass correlations between 0.31 and 0.49 (see Table 2). Measurement error in the dependent variable can limit precision, but does not lead to attenuation bias. We generate change scores by subtracting the standardized

baseline observation score from the end-of-summer standardized score. Teachers are ranked almost identically (correlations above 0.95) when we instead calculate the difference in the raw scores, and then standardize by dividing by the standard deviation of the baseline score.

Our primary estimates of interest come from μ_j, which is a coach random effect and can be thought of as the contribution of individual coaches to teacher outcomes above and beyond variables controlled for in the model. The *j* subscript on μ indicates that the random effect is a random variable, and that we could generate an effectiveness estimate for each coach. We are primarily interested in the underlying distribution of the coach effects and the degree of dispersion. A large degree of dispersion—as indicated by a large SD of the coach effectiveness distribution—suggests that it makes a large difference for teachers' instructional practice in terms of the coach with whom they work. Comparatively, a SD of or close to zero would indicate that there is little heterogeneity in effectiveness across coaches. We do not need to calculate the individual coach effects and their distribution, as our random effects model allows us to generate model-based estimates of the variation in changes in the quality of teachers' classroom practice associated with individual coaches. Model-based estimation via restricted maximum likelihood produces a consistent estimator for the true variance of coach effects (Guarino et al., 2015; Raudenbush & Bryk, 2002). Our random effects model shrinks the coach effects back towards the mean based on the precision of those estimates, driven primarily by the number of teachers per coach (mean = 8.2, SD = 2.5).

In our multilevel model, the unit of analysis is the teacher, and teachers are fully nested within coaches meaning that they work with just one coach. All teachers show up in just one year. Coaches also are fully nested within sites. As described above, in our primary analyses, we include site-year fixed effects and thus absorb all of variation at this additional level. In an exploratory analysis, we replace site-year fixed effects with site or site-year random effects in order to examine variation in teacher outcomes at this level versus the coach level.

Findings

Heterogeneity in Effectiveness Across Coaches

In Table 3, we show estimates of 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. We pool data across all sites and years where a rater other than teachers' coach provided instructional quality scores on the summative, end-of-summer, post-coaching observation. We find that estimates of the coach-level variability are consistent across the four outcome measures: a 1 SD increase in coach effectiveness is associated with a 0.43 SD increase over the course of the summer in the quality of teachers' classroom practice on *Culture of Learning, Essential Content, Demonstration of Learning*, and the composite measure that is an average of the other three scores. In other words, having a coach at the 84th percentile in the distribution of effectiveness relative to a coach at the 50th percentile moves the median-performing teacher to 67th percentile in instructional quality. In Table 3, stars correspond to *p*-values from a likelihood-ratio test that the coach-level variation is different from zero in the multi-level model, relative to a linear model. All four estimates of the coach-level variation are statistically significantly different from zero.

Our estimates of coach effectiveness heterogeneity also can be interpreted relative to average coaching program effects documented in other research. Average program effects are the differences in instructional quality measures for coached versus non-coached teachers. Estimates of coach effectiveness heterogeneity are the difference from average program effects for teachers assigned to a highly effective versus a less effective coach, where the average of the coach effects is the average program effect. In our study, we do not have data to make comparisons between coached and non-coached teachers, and instead focus only on the between-coach comparisons. But, imagine that the average effect of TNTP coaching on measures of instructional practice is similar to other studies at 0.5 SD (Kraft et al., 2018). Pairing this estimate with estimates of coach effectiveness heterogeneity

presented in Table 3, we infer that teachers assigned to a coach 1 SD above the mean in effectiveness would improve in their classroom practice by roughly 0.93 SD relative to a hypothetical comparison group of non-coached teachers; 0.93 SD is the sum of the average program effect (assumed to be 0.5 SD) and the heterogenous effect for having a highly effective coach (a boost of 0.43 SD). In contrast, teachers assigned to a coach 1 SD below the mean in effectiveness would improve by 0.07 SD relative to the hypothetical comparison group; 0.07 SD is the sum of the average program effect (0.5 SD) plus the heterogenous coach effect, which here is negative (-0.43 SD). Teachers assigned to a coach any less effective than that would have no discernible benefit—and potentially a negative effect—relative to a hypothetical set of non-coached teachers.

Identification Checks and Robustness Tests

In part because the main results of this paper can be summarized in a single number (0.43 SD), we conduct several robustness tests to make sure that the estimate is accurate. To begin, we probe the key identifying assumption of this paper: that our value-added model allows us to estimate the true underlying variation in coach effectiveness that is not confounded with the non-random sorting of teachers to coaches.

In Table 4, we show estimates of placebo tests in which we randomly assign teachers to coaches and re-fit our models. This randomization process means that teachers are linked to coaches that they did not actually work with, and so we expect the placebo test to return coach "effect" estimates of zero. Values that are statistically significantly greater than zero would suggest that our main models are falling short of capturing the causal effect of individual coaches. We conduct the randomization process in two ways, randomly assigning teachers to coaches: (i) across all sites and years, and (ii) within site and year. The last level of randomization is the one within which actual coach-teacher matches are made. However, a limited number of coaches per site-year combination means that a randomization process could idiosyncratically pair groups of teachers with their actual coach.

For both randomization procedures, we hold constant coach-teacher ratios. For parsimony, we focus on the composite measure of teaching practice as the outcome. We find that our coach effectiveness model passes the placebo tests. Both estimates of the SD of the coach-level variance are zero.

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 (indicated by "--" in Table 3). To confirm that our placebo estimates are true zeros, we estimated results to 10 decimal places, finding similar results.

Similarly, we conducted a set of falsification tests that estimate the "effect" of coaches on observable background teacher characteristics (i.e., gender, race/ethnicity), controlling for a baseline measure of the outcome, and site-year and certification area fixed effects. Positive and statistically significant estimates 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). Here, we have to include baseline measures of teaching practice on the right- rather than the left-hand side of the equation, given that the outcome measures of the falsification test are teacher-level characteristics. We find that the coach-level variation is zero or very close to zero when predicting most of the teacher demographic dummy variables. When predicting whether or not a teacher is Asian, we observe non-zero variation at the coach level, but the estimate is roughly a tenth as large as when predicting changes in teacher practices. These patterns suggest that our covariates likely have accounted for potential sorting bias. Here too, we show estimates to three decimal places while also confirming that estimates are similar to 10 decimal places.

Next, in Appendix Table 1, we present results when re-estimating coach effects as a set of fixed effects, rather than as random effects. When introducing our model, we proposed a random effects specification that produces a consistent estimator for the true variance of coach effects that also accounts for imprecision in coach effects when estimated from few teachers (Raudenbush &

Bryk, 2002). However, an assumption of this sort of specification is that the random effects—and anything else that shows up in the residual—are uncorrelated with covariates in the fixed portion of our model. In the discussion immediately above, we show evidence in support of this claim. As an additional check, we take an alternative approach that estimates coach effects as a set of fixed effects, which are not assumed to be uncorrelated with covariates (Guarino et al., 2015). Calculating the variation across individual coach fixed effect estimates will bias our variance estimates upward because it conflates true variation with estimation error. To help offset this sort of inflation, we summarize the estimated standard errors of the coach fixed effects to estimate the sampling error variance. Then, we shrink the coach fixed effect estimates by the signal-to-noise ratio.

The SD of the shrunken coach fixed effect estimates (0.5 to 0.55 SD) are, indeed, somewhat larger than the SD of the coach random effects (0.43 SD) but tell a similar story: there is substantial variability in effectiveness across coaches in terms of changes in the quality of teachers' classroom practice. We do conduct formal hypothesis tests that the variation in coach fixed effect estimates is statistically significantly different from zero, as our estimates are simple univariate descriptive statistics (i.e., SDs) of the individual coach fixed effect.

Coaches versus Sites and Generalizability

Finally, we examine the extent to which variation in coach effectiveness is driven by specific sites. Even though all TNTP sites operate under a common coaching model, 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. To examine this hypothesis, we expand the analysis sample to include a census of teachers and coaches that TNTP worked with over a six-period. The larger sample includes 14 unique sites and 40 site-year combinations, facilitating estimation of an additional random effect parameter at this level (compared

to the primary estimation sample with five sites and 11 site-year combinations). Because we rely on the larger sample, we view these analyses as exploratory.

In Appendix Table 2, we start by re-estimating the coach-level variation, but in the expanded sample. We replace site-year fixed effects with year fixed effects because we aim to estimate—rather than absorb—variation across sites. We find that 1 SD in coach effectiveness is associated with a 0.42 SD increase in changes in the composite measure of teachers' classroom practice, which is quite similar to our primary estimate (0.43 SD). Given that we no longer include fixed effects for sites, this estimate of the coach-level variation subsumes site-level variation.

We next nest coaches within sites in our multi-level model and random effects structure. Predictably, the coach-level variation attenuates slightly (0.36 SD) because some of the variation across coaches is now attributed to sites. The site-level variation is similar to the coach-level variation (0.34 SD). When we instead nest coaches within site-year combinations, we get almost identical results. 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-level variation range from 0.26 SD to 0.34 SD. We infer from these results that heterogeneity in effectiveness across individual coaches is not simply a product of the sites within which they work. We also infer that our conclusions regarding coach effectiveness heterogeneity generalize across TNTP sites, coaches, and teachers, despite differences between the full sample and the "rater-not-coach" sample in terms of who rated instruction and some differences in the demographic characteristics of teachers and coaches.

Discussion and Directions for Continued Research

Using a value-added approach similar to the teacher effectiveness literature, we present evidence that individual coaches are the key ingredient for success of instructional coaching programs. Across a range of models and specifications, we observe substantial variation across coaches in how

teachers improve in the quality of their instructional practice. The magnitude of coach-level heterogeneity in effectiveness is particularly large when compared to the average effect of coaching programs. In our primary specification and sample, we find that a 1 SD increase in coach effectiveness is associated with a 0.43 SD increase in multiple dimensions of teaching practice; a 2 SD increase in coach effectiveness is associated with a 0.86 SD increase in teachers' instructional quality. Comparatively, meta-analytic estimates indicate that instructional coaching—on average across programs and across coaches—improves teacher practice by roughly 0.5 SD, which in turn translates into improvements in student test scores of roughly 0.2 SD (Kraft et al., 2018). If we assume that TNTP's coaching model has similar average effects as other programs, then we can infer that assignment to a highly effective coach at least 1 SD above the mean in effectiveness would roughly double the average effect; assignment to a coach 1 SD below the mean or lower would result in no effects of coaching.

On one hand, our study's focus on teachers and coaches working in school districts across the U.S. increases generalizability relative to other similar studies conducted with a small number of coaches or in a single setting. On the other hand, we focus only on the pre-service component of teacher training in an alternative-route certification program, and so we cannot make claims regarding variation in coach effectiveness during in-service professional development nor in other types of training and certification programs. While pre-service teacher coaching has less coverage in the empirical literature base compared to in-service programs, recent experimental evidence of pre-service coaching in a traditional training route identifies effects on teacher practice that are on par with or larger than effects of in-service coaching (Cohen et al., 2020).

To confirm and extend these findings, future research might estimate coach effects under experimental conditions, where coaches are randomly assigned to teachers. Future research might also link coaches 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 from our study are quite large and suggest that these relationships may further translate into changes in student outcomes. These lines of inquiry could be conducted both during pre-service training and in-service development provided to more veteran teachers.

Additional lines of inquiry should explore the specific coach characteristics and coaching techniques that help explain the variability in coach effectiveness that we observe. In other words, what are the key domains of coach characteristics that explain differences in effectiveness, and how can this knowledge be leveraged for recruitment and screening of, and professional learning for coaches? Our study does not address this important practice and policy question directly, though we believe that there is some guidance in the literature that can serve as a bridge between our work and future research. 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 skill have not been linked to changes in teacher practice, particularly in samples that can lead to generalizable conclusions. As such, we advocate for continued research that pairs rich data collection on coaches and their coaching activities with the coach-teacher links and teacher outcome measures that we use in this study.

Implications for Scalability

Ultimately our findings have broader implications for teacher training and development organizations, schools, and districts interested in building or expanding their coaching programs. Currently, school districts spend approximately $18 billion on teacher development programs 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; Harris & Sass, 2011; 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 (Kraft et al., 2018).

Further, the overall costs of coaching programs are comparable to other training and development 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. (All cost estimates are adjusted to 2022 dollars.) Examining coaching in an alternative-route teacher certification context, Kaufman et al. (2020) estimate that coaching comprises roughly a third of total per-teacher costs, at roughly $13,000. The literature on costs of more traditional teacher development and training is older, but suggests that expenditures are similar, at $3,100 to $11,700 per teacher per year (Miles et al., 2004). Given that coaching has similar costs and larger effects than more traditional development offerings, the former is likely to be more cost effective than the latter. Further, because coaching purposefully is individualized and differentiated, it may make 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. In pre-service training contexts such as ours, all teacher trainees likely need coaching, so this proposition would apply more to in-service instructional coaching.

At the same time, adopting and scaling instructional coaching is a risky proposition without knowing how to identify effective coaches—whose salary is the key cost driver of coaching programs (Kaufman et al., 2020; Knight & Skrtic, 2021)—and how to recruit, train, and support more of them. Our findings suggest that highly effective coaches have large impacts on changes in the quality of teachers' classroom practice, while less effective coaches likely return small (if any) benefit for teachers. Within small-scale coaching programs that often operate under best-case conditions, recruiting highly skilled coaches likely is doable and sustainable (Kraft et al., 2018). However, a challenge emerges for larger-scale coaching programs that have to recruit, hire, and train many coaches, and that potentially pull highly effective teachers out of classrooms to serve in these roles. The inherent tradeoff between personnel quantity and quality can be seen from policy decisions in the teacher workforce. For

example, California's decision to reduce class size in the late 1990s necessarily required hiring many more teachers, which resulted in lower qualifications of incoming teachers relative to current teachers (Jepsen & Rivkin, 2002; Stecher et al., 2001). The class size reduction policy did result in improved educational outcomes, but at much smaller magnitudes than documented in prior research.

Results from our study do not directly solve the recruitment challenge described above. Instead, the results serve as a word of caution for school systems: they need to be thoughtful in *who* they recruit to serve in expanding instructional coach roles and *where* these individuals might come from. At the same time, 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. Additional research that examines specific coach characteristics and coaching moves that explain variability in coach effectiveness could also be used to develop screening instruments and coach development offerings.

Rigorous empirical evidence indicates that coaching should be at the forefront of instructional improvement efforts. Scaling 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

Bastian, K. C., Patterson, K. M., & Carpenter, D. (Online 2020). Placed for success: Which teachers benefit from high-quality student teaching placements? *Educational Policy*.

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. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*, *13*(3), 281-309.

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

Blazar, D., & Kraft, M. A. (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.

Britton, L. R., & Anderson, K. A. (2010). Peer coaching and pre-service teachers: Examining an underutilised concept. *Teaching and Teacher Education*, *26*(2), 306-314.

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208-231.

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.

Domina, T., Lewis, R., Agarwal, P., & Hanselman, P. (2015). Professional sense-makers: Instructional specialists in contemporary schooling. *Educational Researcher*, *44*(6), 359-364.

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/

Foote, M. Q., Brantlinger, A., Haydar, H. N., Smith, B., & Gonzalez, L. (2011). Are we supporting teacher success: Insights from an alternative route mathematics teacher certification program for urban public schools. *Education and Urban Society*, *43*(3), 396-425.

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.

Goldhaber, D., Krieg, J., & Theobald, R. (2020). Effective like me? Does having a more productive mentor improve the productivity of mentees? *Labour Economics*, 63.

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.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical Bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics*, *40*(2), 190-222.

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

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7-8), 798-812.

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

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.

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.

Jepsen, C., & Rivkin, S. G. (2002). *Class size reduction, teacher quality, and academic achievement in California public elementary schools*. San Francisco: Public Policy Institute of CA.

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

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Research Paper. MET Project. Bill & Melinda Gates Foundation.

Kaufman, J. H., Master, B. K., Huguet, A., Yoo, P. Y., Faxon-Mills, S., Schulker, D., & Grimm, G. E. (2020). *Growing teachers from within: Implementation, impact, and cost of an alternative teacher preparation program in three urban school districts. Research Report. RR-A256-1*. RAND Corporation. Retrieved from: https://eric.ed.gov/?id=ED609341

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. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1-36.

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.

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.

Matsko, K. K., Ronfeldt, M., Green Nolan, H., Klugman, J., Reininger, M., & Brockman, S. L. (2020). Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education, 71*(1), 41-62.

McEachin, A., Schweig, J. D., Perera, R., & Opper, I. M. (2018). *Validation study of the TNTP Core Teaching Rubric.* RAND. Retrieved from: https://www.rand.org/content/dam/rand/pubs/research_reports/RR2600/RR2623/RAND_RR2623.pdf

Menzes, A., & Maier, A. (2014). Fast Start: Training Better Teachers Faster, with Focus, Practice and Feedback. *TNTP*. Retrieved from: https://files.eric.ed.gov/fulltext/ED559704.pdf

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

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-257.

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.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher, 47*(7), 405-418.

Ronfeldt, M., Goldhaber, D., Cowan, J., Bardelli, E., Johnson, J., & Tien, C. D. (2018). *Identifying promising clinical placements using administrative data: Preliminary results from ISTI placement initiative pilot* (CALDER Working Paper No. 189). National Center for Analysis of Longitudinal Data in Education Research. Retrieved from: https://caldercenter.org/sites/default/files/WP%20189.pdf

Ronfeldt, M., Matsko, K. K., Greene Nolan, H., & Reininger, M. (2020). Three different measures of graduates' instructional readiness and the features of preservice preparation that predict them. *Journal of Teacher Education, 72*(1), 56-71.

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.

Stecher, B., Bohrnstedt, G., Kirst, M., McRobbie, J., & Williams, T. (2001). Class-size reduction in California: A story of hope, promise, and unintended consequences. *Phi Delta Kappan*, *82*(9), 670-674.

TNTP. (2013). *Leap year: Assessing and supporting effective first-year teachers*. Retrieved from: https://tntp.org/assets/documents/TNTP_Leap_Year_2013.pdf

TNTP. (2014). *TNTP Core Teaching Rubric: A tool for conducting Common Core-aligned classroom observations*. Retrieved from: https://tntp.org/publications/view/tntp-core-teaching-rubric-a-tool-for-conducting-classroom-observations

TNTP. (2018). *The opportunity myth: Technical appendix*. Retrieved from: https://files.eric.ed.gov/fulltext/ED590222.pdf

Walsh, K., & Jacobs, S. (2007). *Alternative certification isn't alternative*. Thomas B. Fordham Institute. Retrieved from: https://www.nctq.org/nctq/images/Alternative_Certification_Isnt_Alternative.pdf

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.

Wilson, S. M. (2014). Innovation and the evolving system of US teacher preparation. *Theory into Practice*, *53*(3), 183-195.

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

	Full Sample		Rater-not-Coach Sample	
	Teachers	Coaches	Teachers	Coaches
Demographics				
Female	0.66	0.66	0.61	0.74
Male	0.30	0.24	0.31	0.18
Missing Gender	0.03	0.10	0.08	0.08
Asian	0.03	0.04	0.05	0.03
Black	0.26	0.25	0.26	0.37
Hispanic	0.04	0.04	0.06	0.05
White	0.40	0.52	0.25	0.42
Multiple Races/Ethnicities	0.06	0.04	0.05	0.04
Missing Race/Ethnicity	0.20	0.10	0.33	0.08
Certification Area				
Early Childhood Education	0.07	NA	0.04	NA
Elementary School	0.24	NA	0.39	NA
English Language Arts (ELA)	0.11	NA	0.11	NA
Math	0.08	NA	0.08	NA
Science	0.09	NA	0.07	NA
Social Studies	0.01	NA	0.02	NA
English as a Second Language	0.04	NA	0.02	NA
Special Education	0.15	NA	0.12	NA
Foreign Language	0.01	NA	0.01	NA
Missing Certification Area	0.20	NA	0.14	NA
Coaching Experience with TNTP				
Total yrs.	NA	1.35	NA	1.38
1 yr. Experience	NA	0.74	NA	0.73
2 yrs. Experience	NA	0.20	NA	0.18
3 or more yrs. Experience	NA	0.07	NA	0.09
Persons (*n*)	3,526	317	749	81
Site-Years (*n*)	40		11	
Sites (*n*)	14		5	

Notes: Total years of coaching experience is measured in years. All other statistics are proportions.

Table 2. Descriptive Statistics for Observation Scores

Observation Scores (1 to 3 Scale)	Univariate Statistics				Reliability		
	Last Score		First Score				
	Mean	SD	Mean	SD	Lesson-Level ICC	Teacher-Level Adjusted ICC	Inter-Rater Agreement
Composite	2.51	0.50	2.25	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%

Notes: ICC = intraclass correlation. 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.

Table 3. Standard Deviation of Coach-Level Variation in Changes in Teaching Practice

	Composite	Culture of Learning	Essential Content	Demonstration of Learning
SD of Coach Random Effect	0.433***	0.425***	0.427***	0.428***
	(0.066)	(0.071)	(0.071)	(0.075)
Teachers (*n*)	749	749	749	749
Coaches (*n*)	81	81	81	81

Notes: Each estimate comes from a separate multilevel model of changes in teachers' standardized observations scores from beginning to end of summer, teacher gender and race/ethnicity, certification area fixed effects, and site-year fixed effects. *** $p<0.001$, on likelihood-ratio tests that the coach-level variation component is different from zero in the multi-level model, relative to a linear model.

Table 4. Placebo Test with Teachers Randomly Assigned to Coaches within Different Blocks

	Composite Measure of Teaching Practice	
	Across Sites and Years	Within Site and Year
SD of Coach Random Effect	0.000	0.000
	--	--
Teachers (*n*)	749	749
Coaches (*n*)	81	81

Notes: Each estimate comes from a separate multilevel model of changes in teachers' standardized observation score on the composite measure from beginning to end of summer, teacher gender and race/ethnicity, certification area fixed effects, and fixed effects for the level at which teachers were randomly assigned. Coach-teacher ratios are held constant across all analyses. "–" indicates that the relevant parameter could not be estimated.

Table 5. Falsification Test Predicting Immutable Teacher Characteristics

	Female	Asian	Black	Hispanic	White
SD of Coach Random Effect	0.000	0.038*	0.000	0.000	0.000
	--	(0.017)	--	--	--
Coaches (*n*)	749	749	749	749	749
Teachers (*n*)	81	81	81	81	81

Notes: Each estimate comes from a separate multilevel model of a teacher demographic characteristics dummy 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. * $p<0.05$, on likelihood-ratio tests that the coach-level variation component is different from zero in the multi-level model, relative to a linear model.

Appendix

Appendix Table 1. Standard Deviation of Coach-Level Variation in Changes in Teaching Practice using Coach Fixed Effects Strategy

	Composite	Culture of Learning	Essential Content	Demonstration of Learning
SD of Coach Fixed Effects	0.502	0.548	0.499	0.514
Teachers (*n*)	749	749	749	749
Coaches (*n*)	81	81	81	81

Notes: Each estimate comes from a separate regression model of changes in teachers' standardized observations scores from beginning to end of summer, teacher gender and race/ethnicity, certification area fixed effects, and coach fixed effects. The coach fixed effect estimates are shrunken back towards the mean by multiplying by the signal-to-noise ratio. Tests of statistical significance are not included, as the estimates presented are standard deviations generated from calculation of univariate statistics rather than model-based estimation.

Appendix Table 2. Standard Deviation of Coach-Level Variation on Changes in Teaching Practice, by Site in Full Sample

	Composite Measure of Teaching Practice					
	All Sites	All Sites	Site 1	Site 2	Site 3	Site 4
SD of Site Random Effect	NA	0.337***	NA	NA	NA	NA
		(0.085)				
SD of Coach Random Effect	0.423***	0.355***	0.261***	0.267***	0.341*	0.281*
	(0.028)	(0.027)	(0.046)	(0.058)	(0.112)	(0.101)
Teachers (*n*)	3,526	3,526	873	719	326	399
Coaches (*n*)	317	317	59	47	36	32

Notes: Each estimate comes from a separate multilevel model of changes in teachers' standardized observations scores from beginning to end of summer, teacher gender and race/ethnicity, certification area fixed effects, and year fixed effects. *** $p<0.001$, * $p<0.05$, on likelihood-ratio tests that the coach-level variation component is different from zero in the multi-level model, relative to a linear model.