INSTRUCTIONAL COACHING PERSONNEL AND PROGRAM SCALABILITY

David Blazar
(corresponding author)
Department of Teaching and Learning, Policy and Leadership
University of Maryland College Park
College Park, MD 20740
dblazar@umd.edu

Doug McNamara
Department of Teaching and Learning, Policy and Leadership
University of Maryland College Park
College Park, MD 20740
dmcnama1@umd.edu

Genine Blue
TNTP
New York, NY 10018
genine.blue@tntp.org

Abstract

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 preservice teachers' instructional practice (roughly 0.25 to 0.3 standard deviation from our preferred sample and models). The magnitude of coach effectiveness heterogeneity is quite similar to average coaching program effects on teaching practice identified in other research. Through a set of alternative model specifications and permutation tests, we rule out the possibility that our estimates of coach effectiveness heterogeneity are driven by nonrandom sorting of coaches to teachers, at least on observable characteristics available in our data, as well as the possibility that these estimates are simply statistical noise. These findings suggest that identifying, recruiting, and supporting highly skilled coaches will be key to scaling instructional coaching programs.

https://doi.org/10.1162/edfp_a_00407
© 2023 Association for Education Finance and Policy

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

1. INTRODUCTION

Instructional coaching is an attractive alternative to one-size-fits-all teacher training and professional development. Compared with traditional, workshop-based programs that generally are ineffective (Yoon et al. 2007; Fryer 2017), one-on-one coaching observation and feedback cycles have large effects on teacher practice (0.34 to 0.63 standard deviation [SD], depending on the size of the program) that translate into meaningful impacts for students (roughly 0.1 to 0.3 SD on test scores; Kraft, Blazar, and Hogan 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 United States, with the number of coaches per student roughly doubling between 2000 and 2010 (Domina et al. 2015) and continued growth of programs and hiring of coaching personnel since then (National Center for Education Statistics [NCES] 2016).

Despite considerable 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 and 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 and 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 (Wong and Nicotera 2006; Denton and Hasbrouck 2009; Connor 2017). For instructional coaching to be a viable intervention across teacher training organizations, districts, and schools, it is necessary to identify, recruit, and hire a 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 coaches could undermine efforts to make coaching a primary—if not the primary—teacher training and development tool.

This paper advances the literature on how to improve teachers' instructional practices generally and on coach-based development activities more specifically by investigating the importance of who implements these programs. We ask, *To what degree do individual coaches vary in their effectiveness at improving the quality of teachers' instructional practice?* Variation in effectiveness across coaches can be thought of as heterogeneity in impacts within an instructional coaching program. To answer this question, we draw on the empirical literature on teacher value added (e.g., Guarino, Reckase, and Wooldridge 2015; Koedel, Mihaly, and Rockoff 2015; Bitler et al. 2021), applied to secondary data from TNTP (formerly called The New Teacher Project) in the context of its alternative-route

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

teacher certification program that relies heavily on coaching during the summer/preservice training period.[1]

The collaboration with TNTP is appealing to examine this topic for several reasons. The nature of TNTP's coaching activities within a formal teacher education program means that the organization collects a rich set of data necessary to make decisions regarding provisional certification, including measures of teaching practice scored by trained observers. These instructional quality scores serve as our outcome of interest. Within their data infrastructure, TNTP also collects coach–teacher and observer–teacher links that are necessary for our analyses. Further, because TNTP's programming is implemented across the United States, our analyses greatly increase generalizability and statistical power relative to the few prior quantitative analyses on this topic with five coaches (Blazar and Kraft 2015, 2019). In our primary analysis, we focus on six years of data collected from one training site where external evaluators provided instructional quality scores. Our results are quite similar when we expand the sample to include data from fourteen training sites spread across thirteen states. The full sample is less preferred, however, as the coaches themselves provided instructional quality scores in other sites.

Relatedly, TNTP's programming speaks directly to the practice and policy question at hand regarding scalability. As described by Kraft, Blazar, and Hogan (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 twenty coaches per site and year. Because of the seasonal nature of TNTP's summer training, most of the coaches in our sample worked with the organization in this capacity for just one or two years (though some also worked as coaches or instructional leaders in school districts during the school year). As coaching programs continue to expand—potentially through career ladder programs where coaching is one step in the advancement process (Espinoza et al. 2018)—personnel characteristics may reflect those of TNTP. For example, in a statewide instructional coaching program in Florida, roughly half of the coaches served in this role for one or two years (Marsh, McCombs, and Martorell 2012).

Overall, our value-added models identify substantial variability across coaches in terms of changes in teacher practice. A 1 SD increase in coach effectiveness results in upwards of a 0.3 SD increase in a measure of instructional quality captured on TNTP's observation instrument in the preferred external evaluator sample, as well as upwards of a 0.36 SD increase in the full sample. These effects are quite large relative to average effects of scaled-up instructional programs that are most similar to TNTP's on

1. While the literature base on instructional coaching to date has focused on in-service training and development (Kraft, Blazar, and Hogan 2018), there is growing attention to coaching in preservice training in traditional certification programs (e.g., Britton and Anderson 2010; Cohen et al. 2020) and alternative-route certification models like TNTP's (e.g., Foote et al. 2011; Kaufman et al. 2020). We still refer to preservice 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.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

measures of teachers' classroom practice, as documented in other research (0.34 SD; Kraft, Blazar, and Hogan 2018). In these other studies of larger-scale programs, average coaching effects on teaching quality further translate into meaningful student test-score gains (roughly 0.1 SD), suggesting that teachers' students are likely to feel the impact of highly effective versus less effective coaches as well (though we do not have student-level outcomes in our dataset).

Because the main finding of this paper is captured in a single estimate of coach effectiveness heterogeneity, we engage in a number of robustness tests to convince readers that this estimate is not simply capturing nonrandom sorting of coaches to teachers or statistical noise. Both of these issues are key concerns in the value-added literature on teachers (T. Kane and Staiger 2008; Guarino, Reckase, and Wooldridge 2015; Koedel, Mihaly, and Rockoff 2015; Bitler et al. 2021). We provide evidence of minimal sorting of teachers to coaches and to raters based on observable characteristics, including teachers' gender, race/ethnicity, and, most importantly, teachers' baseline classroom observation score. In turn, estimates of coach effectiveness heterogeneity are very similar when we include versus exclude teacher-level controls, as well as when we simultaneously model rater effects. We also show that our preferred model-based estimates of coach effectiveness heterogeneity from random-effects models are similar to those from coach fixed-effect models that can account for the partial correlations between coach–teacher matches and pretreatment teacher characteristics. Additionally, in a subset of years from the external evaluator sample where school placement data is available, we find that estimates are similar in models that include school fixed effects, which aim to parse the effect of coaches from confounding treatments and resources provided by each school site (e.g., mentor teachers that host preservice teacher candidates in their classrooms).

With regard to noise, we show that estimates are similar when we fit error-in-variables models to adjust for measurement error in these instructional quality scores. Further, we perform a series of permutation tests that randomly assign coaches to teachers and reestimate our value-added models, which provides a benchmark for what coach effects might look like simply due to noise or sampling variation (Bitler et al. 2021). These analyses suggest that there may be some systematic unexplained variation across coaches, but that this variation is too small to negate our primary conclusion about the importance of individual coaches to improving teachers' classroom practice within instructional coaching programs.

2. 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 the scalability of instructional coaching programs.

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

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

differ not only in the quality of their classroom instruction (Bell et al. 2012; T. J. Kane and Staiger 2012; H. C. Hill, Blazar, and Lynch 2015), but also in their subsequent impacts on students' test scores and social-emotional development (Hanushek and Rivkin 2010; Kraft 2019). Our analyses also align with newer lines of research that find substantively meaningful variation across principals (Grissom, Kalogrides, and Loeb 2015) and guidance counselors (Mulhern 2022) 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 and 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 issue is that we must consider not just whether individuals vary 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 are roughly 0.15 SD, and estimates on dimensions of students' social-emotional development often are larger (Nye, Konstantopoulos, and Hedges 2004; T. Kane and Staiger 2008; Blazar 2018; Kraft 2019). 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 and social-emotional gains, the effect of varied educational interventions, and policy-relevant gaps between students from different backgrounds (C. Hill et al. 2008; Soland et al. 2022). 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.

By 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 (Wong and Nicotera 2006; Denton and Hasbrouck 2009).

Exploratory analyses of coach characteristics and practices indicate 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

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

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, McCombs, and Martorell 2012; Yopp et al. 2019; Russell et al. 2020; Shannon et al. 2021). 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 from the impacts they have on desired outcomes. 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 suggest 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 larger sample.

A related line of research, often drawing on large samples in administrative datasets, considers links between cooperating or mentor teachers in preservice field placement settings and mentee teacher outcomes. By hosting preservice 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 (Ronfeldt, Brockman, and Campbell 2018; Goldhaber, Krieg, and Theobald 2019; Bastian, Patterson, and Carpenter 2020), 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; Ronfeldt, Brockman, and Campbell 2018; Matsko et al. 2020; Ronfeldt 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 and Showers 1981). Because instructional coaching—like cooperating or mentor teacher placements—is personnel-focused, we hypothesize that there is some degree of

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

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?

3. THE TNTP COACHING MODEL AND PRESERVICE TRAINING CONTEXT

We explore variation in coach effectiveness in the context of TNTP's instructional coaching model for preservice teachers. TNTP began as an alternative-route teacher certification entity and has trained and certified over fifty thousand 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 and 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, preservice 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 and 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 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 and Showers 1981). On average over the course of TNTP's summer training period, teachers spend at least thirty-two 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

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

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 preservice training and coaching model, see Menzes and Maier (2014).

While TNTP coaching and preservice training operates under a common organizational model, individual coaches are the program implementers and they do so with guidance from site managers. In most 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 teachers who completed training with TNTP 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 forty 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.

4. DATA AND SAMPLE

To answer our research question regarding heterogeneity in effectiveness across instructional coaches, we rely on data collected by TNTP across six years (2014 through 2019) and fourteen summer training sites located across thirteen states. A key feature of the data is that we can directly link coaches to preservice teachers and, then, to performance outcomes vis-à-vis observations of teachers' classroom instruction. In a given year, coaches work with an average of 9.2 teachers (see table 1), spread across several schools that serve as field placements. Descriptive evidence from the data—as well as anecdotal evidence from TNTP—suggest that coach–teacher and school–teacher matches are driven largely by coaches' content expertise and the subject area in which teachers seek certification (see table 1 for a list of all possible certification areas offered by TNTP). We observe that coaches worked with teachers across two certification areas, on average, with instances of multiple certification areas generally having close alignment (e.g., math and science, special education and general elementary, English language arts and English language learners). In the one site where school placement data is available, we also observe that 70 percent of schools hosted teachers seeking certification in just one content area. We do not have information on mentor teacher assignments for any sites or school years.

We define two analysis samples. The first is our preferred sample that includes all six years of data from one summer training site where there was a programmatic decision to have external raters observe and score teachers' instruction both prior to

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Table 1. Characteristics of Teachers and Coaches

	External Evaluator Sample		Full Sample	
	Teachers	Coaches	Teachers	Coaches
Demographics				
Female	0.63	0.66	0.66	0.67
Male	0.34	0.31	0.30	0.21
Missing Gender	0.04	0.03	0.03	0.12
Asian	0.03	0.03	0.03	0.03
Black	0.22	0.28	0.26	0.25
Hispanic	0.02	0.03	0.04	0.04
White	0.46	0.53	0.40	0.52
Multiple races/ethnicities	0.05	0.06	0.06	0.04
Missing race/ethnicity	0.22	0.03	0.20	0.12
Certification area				
Early childhood education	0.01	NA	0.07	NA
Elementary school	0.24	NA	0.24	NA
English language arts (ELA)	0.15	NA	0.11	NA
English as a second language	0.04	NA	0.04	NA
Foreign language	0.02	NA	0.01	NA
Math	0.08	NA	0.08	NA
Science	0.10	NA	0.09	NA
Social studies	0.05	NA	0.01	NA
Special education	0.19	NA	0.15	NA
Missing certification area	0.12	NA	0.20	NA
Coaching experience with TNTP				
1 yr. experience	NA	0.66	NA	0.74
2 yrs. experience	NA	0.25	NA	0.19
3 or more yrs. experience	NA	0.09	NA	0.07
Unique persons	399	32	3,526	317
Person-years	NA	46	NA	430
Teachers per coach (by year/across years)	9.2/12.5		8.2/11.1	
Years	6		6	
Sites	1		14	

Notes: The external evaluator sample restricts to one site where outside raters other than teachers' own coach provided observation scores that serve as the outcome measure. Average years of coaching experience with TNTP is calculated from a coach-level dataset, taking the maximum number of years observed for each coach.

and after all coaching activities. The nature of coaching models organized around observation and feedback cycles led by the coach means that coaches often observe and score teachers' instruction. However, using coach-provided 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. We refer to the preferred sample where external raters rather than coaches provided instructional quality scores as the "external evaluator sample," which includes a total of 399 teachers, 32 unique coaches, and 46 coach-years.[2] To increase generalizability, we extend our analyses to the full

2. The administrative records from TNTP include rater identification numbers and names in summers 2016 through 2019, which we crosscheck against coach names. Our analytic dataset also includes summers 2014 and 2015, when we do not have rater numbers or names. However, for the external evaluator sample/site, we

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

census of sites and years that includes 3,526 teachers, 317 unique coaches, and 430 coach-years. Excluding the external evaluator sample from this full sample, 40 percent of post-coaching observation scores were provided by an external evaluator; 9 percent of baseline scores were provided by an external evaluator. Therefore, we are more cautious about the inferences we can draw from this larger sample.

In table 1, we show that the external evaluator and full samples of teachers are fairly similar to each other: Teachers are roughly two thirds female, one quarter black, and two fifths white. (Roughly 20 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 (Shen 1997; NCES 2016). Demographic characteristics of coaches are similar to those of teachers: Roughly two thirds are female, one quarter are black, and half are white. Across the six years of data we have access to, over 90 percent of coaches—in both the external evaluator and full samples—show up just once or twice; in our preferred external evaluator sample, 66 percent of coaches are observed in the data for one year and 25 percent are observed for two years. Anecdotal evidence from our TNTP partners suggests that many of these individuals held coaching or instructional leadership positions outside of TNTP—such as in their school district jobs during the academic year—though we are not able to observe these roles in the administrative records available for this study.

Our outcome measures capture the quality of teachers' classroom practice as rated on the TNTP-developed rubric (TNTP 2014). As noted above, this rubric guides the coaching process by providing formative assessment data and to identify areas for improvement. Scores captured at the end of the summer training period also are used for summative decisions regarding provisional certification. Trained evaluators—including coaches and the external evaluators hired specifically for scoring instruction—observed and rated teachers' instruction up to six times over the course of the summer, though our analyses focus on scores from two time points: The first observation conducted before the start of coaching and the last observation conducted after coaching. 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 forty to fifty hours a year of observation practice.

While the observation rubric includes three dimensions of instructional practice, we focus on one—*Demonstration of Learning*—that is least prone to ceiling effects that can arise from the 3-point rating scale from 1 (Ineffective) to 3 (Developing). From a practice-based standpoint, the rating scale is purposefully truncated given the nature of the preservice training period, where preservice teachers are unlikely to become effective or highly effective in their first couple of months in the classroom. *Demonstration of Learning*—which asks whether all students demonstrate that they are learning, and focuses on the extent to which teachers check for student understanding and respond

received confirmation from the site director that external raters, rather than coaches, scored teachers' instruction in all years 2014 through 2019.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Table 2. Descriptive Statistics on *Demonstration of Learning* Observation Scores (1 to 3 Scale)

	External Evaluator Sample			Full Sample		
	Baseline	Outcome	Change	Baseline	Outcome	Change
Univariate statistics						
Mean	1.84	2.39	0.56	1.97	2.31	0.33
SD	0.79	0.73	0.96	0.71	0.70	0.84
Skew	0.30	–0.76	–0.12	0.04	–0.50	–0.01
Intra-Class Correlations (ICC)						
Lesson			0.44			
Teacher			0.58			
Rater			0.13			

Notes: ICCs are calculated from all available lessons where a rater other than the coach scored instruction so that estimates are not confounded by coaches' work with teachers. We do not disaggregate ICCs by baseline or outcome, as multiple lessons per teacher are needed to estimate and differentiate between lesson- and teacher-level ICCs. Following a generalizability framework, teacher-level ICCs are adjusted for the modal number of lessons per teacher.

to student misunderstandings—appears to be the most difficult of the teaching tasks on the TNTP rubric, leaving room for improvement over the course of the coaching period. The average baseline/pre-coaching score on this dimension for the preferred external evaluator sample is below the scale's midpoint (baseline mean 1.89; SD 0.8), with an average pre/post gain across teachers of 0.56 scale points. Further, the baseline and change scores have minimal skew (see table 2). Nonetheless, the 3-point rating scale does present the possibility of ceiling effects and right censoring,[3] a topic we return to below. In supplemental analyses, available in a separate online appendix that can be accessed on *Education Finance and Policy*'s website at https://doi.org/10.1162/edfp_a_00407, we consider coach effects on the two other dimensions of teaching practice from the TNTP rubric, but where there is much less room for improvement and a greater likelihood of right censoring. We standardized observation scores to have a mean of 0 and SD of 1.

Scores generated from the TNTP rubric 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 our dataset, we also provide evidence that observation scores capture the underlying construct of interest—that is, the quality of classroom instruction—by calculating intraclass correlations (ICCs). The ICCs capture the proportion of total variation in observation scores that come from construct-relevant sources of variation (i.e., lessons, teachers) as opposed to construct-irrelevant sources of variation (i.e., raters). We use all available lessons—prior to and at the end of coaching—where a rater other than the coach scored instruction so that estimates are not confounded by the effect of individual coaches on teachers' practice. The lesson-level ICC for *Demonstration of Learning* is 0.44, which is similar to other studies in which trained observers

3. In online appendix table 1, we show bivariate distributions of baseline and outcome scores on *Demonstration of Learning* to examine the extent of possible ceiling effects and right censoring. For example, of those teachers in the preferred external evaluator sample that earned a baseline score of 3 (the top score), 71 percent scored a 3 at the end of coaching; teachers who scored 3 on both the baseline and post-coaching observations make up 17 percent of this sample.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

scored teachers' instruction (Bell et al. 2012; H. C. Hill et al. 2012). While our analyses focus on changes in these lesson-level scores, we also note that the teacher-level ICC that accumulates information across lessons is higher (0.58). Following a generalizability framework, we adjust teacher-level ICCs for the modal number of lessons per teacher. In comparison, we identify more modest variation at the rater level (0.13), which also is similar to studies where trained researchers and practitioners scored the quality of teachers' classroom practice (H. C. Hill et al. 2012). Nonzero variation at the rater level captures harshness/leniency across raters and will introduce some noise into estimates of coach effectiveness heterogeneity. Rater-level variation could also introduce bias to our estimates of coach effectiveness heterogeneity, if this variation were correlated with teacher or coach assignments. However, evidence from sorting and robustness tests presented below indicate that rater-level variation is unlikely to drive our results related to coach effectiveness heterogeneity.

5. 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 value-added model of the following form:

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

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. Our primary estimates of interest come from the standard deviation of the coach effects, μ_j, which is a measure of heterogeneity in coach effectiveness. This can be thought of as the contribution of individual coaches to teacher outcomes above and beyond variables controlled for in the model (i.e., gender, race/ethnicity, and certification area), all included in the vector $I_{j(t-1)}$. We further condition on site-year fixed effects, δ_{st}, which is the level at which coach–teacher matches are made. In the preferred external evaluator sample with just one site, we use year fixed effects. According to TNTP and the data, expertise in specific certification areas, site, and year appear to be the primary factors considered when matching coaches to teachers. In a robustness test, we add school fixed effects to our model in order to account for other inputs at the school level (e.g., mentor teachers) that may be correlated with coach assignment. However, teacher–school links only are available from the one site where external evaluators provided classroom observation scores of teachers' instruction, and from a subset of years in this sample.

There are two primary concerns when interpreting estimates of coach effectiveness heterogeneity from equation 1: (i) nonrandom sorting of coaches to teachers, and (ii) statistical noise. Because coaches are not randomly assigned to teachers, there may be variables omitted from our model related both to the coach with whom a teacher works and to teachers' classroom observation score (e.g., coach-teacher personality match), which would bias our results (T. Kane and Staiger 2008). Even if the observable characteristics that we include in our model sufficiently account for nonrandom sorting, it is possible that a large share of the variability in our estimates of coach effectiveness is attributable to factors that have nothing to do with the coach themselves, such as sampling error and measurement error in the teacher observation scores (McCaffrey et al. 2009; Schochet and Chiang 2013). Following guidance on the

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

literature on how best to address these two concerns, we explore a range of alternative specifications, including random versus fixed effects models, gain-score versus lagged-score models, and additional models that aim to account for measurement error and rater effects in the classroom observation scores (Koedel, Mihaly, and Rockoff 2015).

We preference random-effects models primarily for model-based estimation of variation in coach effects that employ shrinkage. Model-based estimation via maximum likelihood produces a consistent estimator for the true variance of coach effects (Raudenbush and Bryk 2002; Guarino, Reckase, and Wooldridge 2015). Our random effects model shrinks the coach effects back toward the mean based on the precision of those estimates, driven primarily by the number of teachers per coach. In our data, the number of teachers per coach—across all years—is moderate (mean 12.5 in the external evaluator sample and 11.1 in the full sample; see table 1), meaning that shrinkage is particularly important.[4] In comparison, value-added models of teacher effects on student outcomes generally are fit to data with roughly twenty-five to thirty students per class. Because most of the coaches in our sample have limited experience with TNTP (see table 1), our analyses estimate coach effectiveness heterogeneity pooling across all available years of data. We also note that the total number of unique coaches in our external evaluator sample is moderate ($N = 32$). However, there is some discussion in the literature that a minimum of five groups is sufficient to estimate variance components (Harrison et al. 2018), and there are established procedures for fitting multilevel models with small sample sizes including restricted maximum likelihood (as opposed to full maximum likelihood; Harville 1977), which we use in our analyses. Tests of model fit as well as permutation tests, described below, provide evidence that we are able to detect true effects of coach effectiveness heterogeneity even in a limited sample. 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.

One concern with random-effects models is that they assume that variables in the fixed portion of the model (e.g., background teacher characteristics) are uncorrelated with μ_j, which may not be the case without random assignment of coaches to teachers (Guarino, Reckase, and Wooldridge 2015). We formally test for sorting on observed teacher characteristics in a set of robustness tests described below (finding minimal evidence of sorting). We further explore coach fixed-effects models that allow for correlation between pretreatment teacher characteristics and the set of coach fixed effects. Here, we predict the individual coach effects and then calculate the SD in these predicted estimates across coaches. The SD of coach fixed effects always is larger than the model-based estimate from the random-effects specification because it does not apply a shrinkage factor. We can manually estimate and apply a shrinkage factor as the signal-to-noise ratio in the coach fixed-effect estimates:

4. We do not make any restrictions on the minimum number of teachers per coach in our analytic sample because estimators that use shrinkage downweight the influence of coaches with few teachers. In the external evaluator sample, one coach (out of thirty-two) worked with one teacher and the rest worked with at least five. In the full sample, twenty-six more coaches (out of 285) worked with fewer than five teachers. When we exclude these coaches from our analyses, estimates of coach effectiveness heterogeneity are very similar, across both samples and all model specifications.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

$$\gamma_j = \frac{\sigma_\mu^2}{\sigma_\mu^2 + (\sigma_\varepsilon^2 / n_j)}, \tag{2}$$

where σ_μ^2 is the variance in the coach fixed effect estimates (i.e., the signal), σ_ε^2 is the model error variance, and n_j is the number of teachers per coach. At the same time, this approach to manual shrinkage is limited in our data given that most coaches are observed for just one year and with a moderate number of teachers per year (table 1), whereas more years of data generally are needed to accurately calculate the correct shrinkage factor (T. J. Kane, Rockoff, and Staiger 2008; Bitler et al. 2021).

For both random- and fixed-effects models, we estimate coach effectiveness heterogeneity using gain-score models (where the outcome of interest is a change score, as in equation 1) and lagged-score models (where the baseline measure is included on the right-hand side of the equation, in the vector $I_{j(t-1)}$). Gain-score models are appealing for two reasons. First, because the baseline and end-of-summer observation scores are on the same scale—and a primary purpose of repeated observation is to capture growth over time—we can simply subtract the two.[5] In contrast, in teacher value-added models of student test-score growth, test scores captured in different grades and school years often are not on vertically equated or comparable scales (M. T. Kane 2017). Second, gain-score models do not introduce bias resulting from measurement error in the baseline score. When a prior measure of the outcome—or other predictors—are measured with error, estimated slopes are attenuated (Lockwood and McCaffrey 2014), further affecting variance components of higher-level units in a multilevel, random-effects model (Ferrão and Goldstein 2009). Because observation scores aimed at capturing the quality of teachers' instruction generally are measured with more error than student test scores, the effect of measurement error on inferences regarding coach effectiveness heterogeneity may be more substantial in our context compared to estimation of teacher value added to student test-score growth. Indeed, the lesson-level ICC for our primary outcome is 0.44, compared to student test score reliabilities that often are in the range of 0.8 to 0.9 (M. T. Kane 2017). There are known procedures for accounting for measurement error in predictors in value-added contexts, including use of error-in variables regression models (Andrabi et al. 2011; Lockwood and McCaffrey 2014). However, these procedures are complicated by heteroskedastic measurement error. And corrections that are applied to multilevel data are quite computationally challenging (Goldstein, Kounali, and Robinson 2008; Battauz, Bellio, and Gori 2011; Lockwood and McCaffrey 2014), moving beyond the scope of this paper.

One tradeoff for gain-score models, relative to lagged-score models, is that they assume no regression to the mean. In other words, the coefficient relating the baseline and end-of-coaching observation scores from a lagged-score specification is assumed to be 1. However, we know this is not true: The correlation between baseline and outcome *Demonstration of Learning* observation score is 0.7, when disattenuated by its

5. In our analyses, we standardize the baseline and outcome observation scores separately, and then subtract the standardized baseline score from the standardized outcome score. We also consider a change score that first subtracts the baseline score from the outcome in their original units, and then divides by the SD of the baseline score. These two change scores are quite similar for *Demonstration of Learning*, given that the baseline and outcome scores have similar SDs in both the external evaluator and full samples (see table 2). Estimates of coach effectiveness heterogeneity also are quite similar using the two change score measures.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

reliability. Because of this incorrect assumption, some scholars describe gain-score models as performing worse than lagged-score models. Simulations show that the latter return "true" teacher effects more frequently than the former (Guarino, Reckase, and Wooldridge 2015). At the same time, the differences are not large. For example, Kane and Staiger (2008) find that lagged-score and gain-score models of teacher effects identify similar magnitudes of teacher-level heterogeneity, which is the statistic of interest in our analysis of coach effectiveness heterogeneity.

Because gain-score models introduce different assumptions than lagged-score models, we implement both in our analyses. We also take advantage of flexibility in lagged-score specifications. Across lagged-score models, we are able to explicitly account for the possibility of nonrandom sorting as a function of teachers' lagged score. Then, we can compare estimates from this approach to a model where we remove the lagged score altogether and reestimate heterogeneity in coach effectiveness. If nonrandom sorting of teachers to coaches based on the lagged score is not severe—or is not present at all—our estimates of coach effectiveness heterogeneity should be quite similar in models that include versus exclude the lagged score. In models that exclude the lagged score, we further explore a specification that crosses coach random effects with rater random effects.[6] This approach allows us to parse the contribution of coaches to teacher outcomes from the influence of raters who provide these scores. While external raters also provide the baseline scores in the external evaluator sample, almost all teachers had different raters at baseline versus after coaching. Therefore, we only focus on modeling rater effects on the outcome and when excluding the baseline score. In the full sample, we estimate the crossed rater-coach random-effects model in instances where an external rater provided the post-coaching score. Similarly, in the coach fixed-effects specification, we specify errors-in-variables models that adjust the variance-covariance matrices based on the reliability of the lagged observation score. As noted above, this approach only is computationally reasonable in the fixed-effects specification.

A final consideration in the estimation of coach effectiveness heterogeneity relates to ceiling effects in the teacher observation scores, which can introduce noise due to censored data as well as bias if right-censored teachers are not randomly sorted across teachers. Koedel and Betts (2010) examine the effects of right-censoring in student test-score data when estimating teacher value added, finding that the estimated teacher effects are only minimally influenced over a wide range of ceiling severity. For example, even when a ceiling is set at the 33rd percentile in the distribution of student performance, with a skew of −2, the magnitude of teacher effects is almost identical to the magnitude of teacher effects from data without any censoring; additionally, the correlation between the individual teacher effect estimates with and without censored data is roughly 0.8. In our data, the skew on teachers' outcome score on *Demonstration of Learning* is less severe (–0.76 and −0.5 in the external evaluator and full samples, respectively).[7] Therefore, below, we focus our identification checks on whether or not

6. External raters are not fully nested nor crossed with coaches: We observe six raters per coach and eight coaches per rater, on average. Crossed rater and coach random effects estimate rater-by-coach effects, in addition to coach effects.

7. In their exploration of the implications of censoring in value-added models, Koedel and Betts (2010) discuss but eventually rule out the benefits of tobit models. They argue that censored-data models to calculate value added are not practically useful because the exact censoring points are unknown, and misidentifying censoring points

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Table 3. Standard Deviation of Coach Effects on Teachers' *Demonstration of Learning* Observation Scores

Model Specifications	External Evaluator Sample	Full Sample
Random-effects models		
Gain score	0.244	0.364
No control for baseline score	0.282	0.237
No control for baseline score, rater effect	0.312	0.212
Lagged score	0.271	0.226
Fixed-effects models (with post hoc shrinkage)		
Gain score	0.400	0.330
No control for baseline score	0.341	0.272
Lagged score	0.339	0.251
Lagged score, measurement error adjustment	0.344	0.299
Teachers	399	3,526
Coaches	32	317

Notes: For random-effects models, likelihood-ratio tests indicate that the coach-level variation component is different from zero in the multilevel model relative to a linear model ($p < 0.05$) in all instances except one: the gain-score model for the external evaluator sample ($p = 0.092$). For all fixed-effects models and in both samples, F-tests on the joint test of significance of the coach dummies indicate statistically significant variation in outcomes across coaches ($p < 0.05$). See table 5 for results from permutation tests that also inform hypothesis testing. The sample sizes for the random-effects models that include crossed rater-by-coach effects are smaller ($N = 21$ coaches and 268 teachers in external evaluator sample, and 102 coaches and 894 teachers in the full sample) because rater identification numbers only are available in a subset of years. Further, in the full sample, we only include observations where an external rater other than the coach provided the outcome observation score.

censored teachers are systematically sorted across coaches (which we find that they are not).

6. FINDINGS

Heterogeneity in Effectiveness across Coaches

We present our main results in table 3, where coefficients describe the SD of coach effects on changes in teachers' instructional practice. More specifically, estimates can be interpreted as the SD increase in the quality of teachers' practice on *Demonstration of Learning* resulting from a 1 SD increase in coach effectiveness. Estimates of or close to zero would indicate that there is little heterogeneity in effectiveness across coaches. As a practical matter for coaching organizations and school districts, an estimate of zero would mean that it does not make a difference which coach that a teacher is assigned; a large degree of dispersion suggests that it makes a large difference for teachers' instructional practice in terms of the coach with whom they work. We specify eight total models, with a combination of coach random- versus fixed-effects specifications, gain- versus lagged-score specifications, models with crossed rater and coach random effects, and error-in-variables models in the coach fixed-effects specification. We focus on interpreting results from our preferred sample where external evaluators provided teacher

produces substantially biased estimates of model parameters. Further, in the value-added context, there may be censoring in both the outcome measure and a lagged measure of performance, and technical solutions for addressing both are highly sensitive to assumptions of the joint distribution of the set of independent variables. Therefore, we do not fit tobit models in our analyses.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

observation scores at baseline and at the end of coaching, though we note that estimates from the full sample are quite similar. This suggests that coaches serving as raters does not markedly alter results. Similarity in estimates across samples also increases generalizability.

Starting first with our preferred random-effects specifications, we find that a 1 SD increase in coach effectiveness results in a 0.24 to 0.31 SD increase in teachers' observed quality of instruction on the *Demonstration of Learning* measure from TNTP's observation rubric. The smallest estimate is from the gain-score model. The estimate from the lagged-score model (0.27 SD) is very close to the estimate from the model that excludes the lagged score (0.28 SD), suggesting that teachers' baseline performance on this measure is unrelated to coach–teacher assignments. We also find a similar estimate of coach effectiveness heterogeneity in crossed models where we simultaneously model coach and rater-by-coach effects (0.31 SD), indicating that variation in outcome scores across raters is not driving our results. Likelihood-ratio tests indicate that the coach-level variation is different from zero in the multilevel model relative to a linear model ($p < 0.05$). The only exception is for the gain-score model for the external evaluator sample ($p = 0.092$). Below, we conduct a set of permutation tests that also inform hypothesis testing against a null hypothesis of no coaching effects.

Estimates of coach effectiveness heterogeneity are quite similar when generated from the set of fixed-effects models that apply post hoc shrinkage. Across models, the SD of coach effects range from 0.34 to 0.4 SD, with the largest estimate coming from the gain-score model. Here, too, inclusion versus exclusion of the baseline observation score as an independent variable in the coach fixed-effect model has a negligible effect on the magnitude of the coach-level variation. For all fixed-effects models, F-tests on the joint test of significance of the coach dummies indicate statistically significant variation in outcomes across coaches ($p < 0.05$). In online appendix figure 1, we show the distribution of the predicted coach effect estimates across all eight models, for both the external evaluator and full samples. Distributions appear roughly normal, and are smoothed out in the full sample with many more coaches and teachers.

Although the primary purpose of this paper is to estimate the variance components rather than individual coach effects, in online appendix table 2 we show correlations between the predicted coach effect estimates across models as a means of understanding stability in rank ordering. We do not adjust these correlations for the reliability of the coach effect estimates because the correlations generally are quite high and adjustments often lead to correlations greater than 1. In the external evaluator sample, correlations between similar random- and fixed-effects specifications are all at or above 0.8 (e.g., correlation between coach random- versus fixed-effect estimates from the lagged-score model is 0.85). Correlations between coach effect estimates from gain-score and lagged-score models also hover around 0.8 (e.g., 0.77 for estimates generated from the gain-score and lagged-score random-effects models). Correlations are very close to 1 for estimates generated from lagged-score models versus models that do not account for the baseline score at all. Adding crossed rater-by-coach random effects does not change rankings of coach effects, relative to a similar random-effects model that excludes a rater effect ($r = 0.98$). For the coach fixed-effects specification, adjustments for measurement error in the lagged score also has a minimal effect on teacher rankings, with correlations above 0.9 for the three other fixed-effects models. A high degree of

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

alignment is similar to findings from the teacher value-added literature (Guarino, Reckase, and Wooldridge 2015). Correlations in coach effect estimates in the full sample generally follow similar patterns.[8]

The findings presented thus far show that the magnitude of the coach-level variation in changes in teachers' instructional practice is fairly consistent across models and samples. But, is the variation large and relevant for policy and practice? One way to interpret estimates in SD units is to convert them to percentile rank. Having a coach at the 84th percentile in the distribution of effectiveness relative to a coach at the 50th percentile (i.e., a 1 SD increase) moves the median-performing teacher to roughly the 60th to the 65th percentile in instructional quality (depending on the estimate used). These estimates are similar to the magnitude of teacher effects on student outcomes (Hanushek and Rivkin 2010). Our estimates of coach effectiveness heterogeneity also can be interpreted relative to average coaching program effects documented in other research: 0.34 SD for scaled-up programs enrolling more than one hundred teachers (Kraft, Blazar, and Hogan 2018). Average coaching program effects capture differences in instructional quality measures for coached versus non-coached teachers, while our estimates of coach effectiveness heterogeneity are the difference in average program effects for teachers assigned to a highly effective versus a less effective coach. We cannot estimate average coaching program effects, as we do not have a comparison group of non-coached teachers.

Are Estimates Driven by Nonrandom Sorting or by Noise?

We conduct a variety of sensitivity analyses and robustness tests in order to examine whether our estimates of coach effectiveness heterogeneity are driven by bias due to nonrandom sorting of coaches to teachers or by statistical noise.

Sorting Tests

The ideal analysis to test for bias due to nonrandom sorting of coaches to teachers is to estimate and compare coach effect estimates under experimental and nonexperimental conditions. The experimental estimates aim to account for all factors that could

8. Another way to examine stability of coach effects is to estimate correlations across years. However, this strategy is challenging in our context due to data limitations. Our dataset includes few coaches with multiple years of data ($N = 11$ in the preferred external evaluator sample, and 84 across the full sample), and few teachers per coach in any given year (average of roughly 9 and maximum of 14). Nonetheless, for the subset of coaches across the full sample with more than one year of data ($N = 84$), we estimate year-to-year correlations of 0.1 to 0.17 across random- versus fixed-effect and gain-score versus lagged-score specifications. For a further subset of coaches with three or more years of data ($N = 23$), we estimate year-to-out-of-year correlations in coach effects between 0.2 and 0.27. At face value, these correlations are weaker than in the teacher value-added literature. For example, McCaffrey et al. (2009) estimate year-to-year correlations in teacher effects between 0.16 and 0.67, depending on the model specification, grade level, and school district. A likely explanation for differences in findings related to intertemporal variability between this study and ours is related to reliability. Teacher observation scores are measured with more error than student test scores, which may attenuate year-to-year correlations in lagged-score models. Further, McCaffrey et al. exclude all teachers with fewer than fifteen students, noting that "estimates for teachers based on very few students will tend to be extremely unstable over time" (2009, p. 587). We cannot make this same restriction because all coaches have fewer than fifteen teachers per year. However, when we limit the sample to coaches with more teachers per year over multiple years ($N = 23$), year-to-year correlations rise to roughly 0.25. We do not have sufficient sample size to estimate year-to-out-of-year correlations here. While weak year-to-year correlations in coach effects highlight sampling and measurement error, permutation tests indicate that we are not simply capturing noise.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Table 4. Coach and Rater Sorting Tests

	Within- versus Between-Coach Variation		
Baseline Teacher Characteristics	External Evaluator Sample	Full Sample	Within- versus Between-Rater Variation
Baseline *Demonstration of Learning* observation score	0.017	NA	NA
Top observation score (of 3) at baseline	0.034	NA	NA
Female	0.000	0.020	0.005
Asian	0.018	0.006	0.026
Black	0.000	0.000	0.008
Hispanic	0.000	0.013	0.021
White	0.000	0.000	0.008

Notes: All estimates are Intra-Class Correlations (ICC) that estimate the percent of total variation that lies between coaches or raters, as opposed to within coaches or raters. ICCs are calculated from variance components generated from random-effects models that condition on teacher certification area and site-year fixed effects; multilevel probit models are used when teacher background measures are dichotomous. We exclude ICCs for baseline observation score measures for coaches in the full sample and for raters, as these individuals provided the baseline observation scores; as such, the ICCs reflect harshness/leniency across raters (see table 2) rather than sorting.

potentially drive coach–teacher matches, including those that generally are observed in quantitative datasets (e.g., demographics, lagged performance) and other possible factors that are generally unobserved (e.g., coach–teacher personality match). A correlation of 1 between experimental and nonexperimental estimates is indication of no sorting bias, and scholars have documented this fact when examining teacher effects on students' math test scores (for a review and meta-analysis, see Bacher-Hicks et al. 2019). Without random assignment of coaches to teachers, we cannot set up this sort of test for sorting on unobservables. However, we can test for bias due to nonrandom sorting on observables in other ways. One test comes from table 3 and online appendix table 2, where we show that estimates of heterogeneity in coach effectiveness and correlations between the individual coach effects are quite similar when we include versus exclude the lagged observation score. If teachers were sorted to coaches nonrandomly based on the lagged score, excluding the lagged score from our model should impact our estimates and rank ordering of coaches. Similarly, inclusion versus exclusion of crossed rater-by-coach effects does not alter results, suggesting that rater-level variation in observation scores (see table 2) is uncorrelated with teacher and coach assignments.

In table 4, we further test for sorting on observables by specifying our preferred random-effects model and replacing the end-of-coaching teacher observation score with baseline teacher characteristics including baseline score on *Demonstration of Learning,* and gender and race/ethnicity dummies. We take a similar approach to examine whether external raters are sorted nonrandomly to teachers, as nonrandom sorting of raters to teachers' lessons also could introduce bias into our estimates of coach effectiveness heterogeneity. Tests of rater–teacher sorting only provide new information above coach–teacher sorting tests in the external evaluator sample, where raters are not coaches. We present estimates as ICCs on a 0 to 1 scale, as the baseline observation score and demographic characteristics are on different scales (i.e., SD units versus binary indicators) and so the variance components themselves are not directly comparable to each other. The ICCs tell us how much of the variation in a given baseline teacher characteristic lies within versus between coaches (or raters). A large degree of

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

between-coach or between-rater variation suggests that coaches or raters are assigned nonrandomly to teachers based on that specific characteristic.

We estimate ICCs that often are close to zero and no higher than 0.034. For example, in the external evaluator sample, we find that 1.8 percent of the variation in the percent of teachers who are Asian lies across coaches, whereas 98.2 percent lies within coaches; and that 2.6 percent of the variation in Asian teachers lies across rather than within raters. We also observe that 1.7 percent of the variation in teachers' baseline *Demonstration of Learning* observation score lies across coaches, whereas 98.3 percent lies within coaches. An ICC for the baseline score of 0.017 translates into 0.13 SD for a standardized observation score. We also calculate the ICC for teachers earning the top baseline observation score (3) on *Demonstration of Learning* in order to examine whether our estimates of coach effectiveness heterogeneity are capturing systematic variation due to nonrandom assignment of coaches to teachers who have little room to improve on a 1 to 3 scale. We find that this is not the case (ICC = 0.03). We do not calculate ICCs for the baseline scores for raters or for coaches in the full sample, as these individuals provided the baseline scores and so nonzero ICCs reflect harshness/leniency in scoring as well as sorting. As another sorting test, we examine whether background characteristics of coaches (i.e., gender, race/ethnicity, years of coaching experience with TNTP) are related to the baseline observation scores of teachers, which we find is not the case (p-value on joint test of significance = 0.229 and 0.545 in the external evaluator and full samples, respectively; not shown in table 4). Pairing the ICCs with estimates from models that include versus exclude the baseline score and rater effects, we conclude that there is some degree of nonrandom sorting but that this does not explain away the key finding of the paper: that coaches differ substantially on their impacts on teacher practice.

This finding is fairly different from the teacher value-added literature, where we know that there is substantial sorting of students to teachers on academic achievement and other factors (Clotfelter, Ladd, and Vigdor 2007). Failing to account for students' baseline achievement can make a large difference in estimates of the effect of teachers on students' current outcomes (Koedel, Mihaly, and Rockoff 2015). A likely reason for this difference relative to our study is that coach–teacher matches appear to be driven largely by logistical rather than substantive reasons—namely, coaches' content expertise and the subject area in which preservice teachers are seeking certification. Another reason is that TNTP site leads have limited information on preservice teachers before they start the program—restricted to information in teachers' application to the program—as well as limited information on seasonal coaches. It may be that coach–teacher sorting is more pronounced during the school year when school leaders have more information on full-time teachers and coaches in order to make strategic pairings. However, this possibility has no bearing on the estimates presented in this paper.

Confounding Treatments within School Placement Sites

Another potential source of bias is confounding treatments, where other inputs such as resources that each school site provides are correlated with coach assignments. One such resource is the mentor teacher with whom a teacher is paired and whose classroom they work in throughout the summer school period. While we do not have information on mentor teachers, we do have some information on the school placements

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

where teachers worked. These data are only available systematically for the one site that contributes to our external evaluator sample, and for four out of the six years ($N = 268$ teachers and twenty-one unique coaches). In this sample, coaches worked with teachers across 6.3 schools, and each school hosted 1.8 teachers and 1.6 coaches per year, on average. In other words, coaches were fairly spread out across schools, while schools were fairly narrow in terms of number of teachers they hosted.

If the coach-level variation described thus far were attributable to variation in resources and mentor teacher effectiveness across school placement sites, then we would expect the SD of coach effects estimated from models that include school fixed effects to be smaller than those that exclude school fixed effects or fall to zero. The coach-level variation—which may be correlated with school-level variation—would be accounted for with the set of school fixed effects. However, this is not the case. In online appendix table 3, we show that estimates of coach effectiveness heterogeneity are slightly larger in models that include school fixed effects (0.36 to 0.39 SD), relative to models that exclude school fixed effects (0.23 to 0.34 SD). From these analyses, we infer that school sites and the resources each provides are not unlikely to drive our estimates of coach effectiveness heterogeneity.

Permutation Tests

Next, we conduct a series of permutation tests in which we randomly assign coaches to teachers and refit our models. According to Bitler et al. (2021), this "approach eliminates any potential for sorting, peer effects, systematic measurement error, and/or true effects; and provides a benchmark for what [coach] 'effects' look like simply due to noise or sampling variation" (p. 902). In other words, the permutation tests provide a benchmark for how much estimation error may mechanically inflate our estimates of the coach effects shown in table 3. In the spirit of randomization inference, by randomizing coaches to teachers we impose a null hypothesis of no true coach effect. We reject this null hypothesis when our estimated coach effects are greater than the 95th percentile of the permuted "effects."

We conduct the permutation process in four ways, randomly assigning coaches to teachers: (i) across all sites and years; (ii) across sites but within year; (iii) within sites but across years; and (iv) within site and year. The last level of randomization is the one within which actual coach–teacher matches are made. However, in instances where there are a limited number of coaches per site–year combination, the randomization process could still pair teachers with their actual coach. For the external evaluator sample that includes just one site, permutations (i) and (ii) are irrelevant. In all randomization procedures, we hold constant coach–teacher ratios. Using our preferred random-effects specification, we run each of the four randomization processes five hundred times for the external evaluator and full samples. We capture statistics from the resulting distribution of estimates from random-effects gain-score and lagged-score models. Because the coach "effects" are bounded on the lower end at zero, the distribution naturally has a long right tail, and the median is lower than the mean. Therefore, we report both, as well as the 95th percentile.

In table 5, we show that the median estimates across permutations are zero in all cases, across levels of randomization blocking (e.g., across versus within sites and years), gain-score and lagged-score random-effects models, and external evaluator and

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Table 5. Permutation Tests of Coach Random "Effects" After Re-Randomization Five Hundred Times

	External Evaluator Sample			Full Sample		
Randomization Blocks	**Median**	**Mean**	**95th Perc.**	**Median**	**Mean**	**95th Perc.**
			Gain-Score Model			
Across sites and years	NA	NA	NA	0.000	0.038	0.126
Within year, across sites	NA	NA	NA	0.000	0.035	0.117
Within site, across years	0.000	0.058	0.234	0.000	0.021	0.099
Within site and year	0.000	0.038	0.188	0.000	0.010	0.080
			Lagged-Score Model			
Across sites and years	NA	NA	NA	0.000	0.026	0.094
Within year, across sites	NA	NA	NA	0.000	0.025	0.093
Within site, across years	0.000	0.049	0.191	0.000	0.017	0.083
Within site and year	0.000	0.034	0.151	0.000	0.006	0.054

Notes: Coefficients presented are SDs of estimated coach "effects" across five hundred randomizations. Because the external evaluator sample consists of one site, the first two randomization blocks are duplicative with the latter two.

full samples. The mean values across five hundred runs are expectedly larger than the median values but never above 0.059. Further, the 95th percentile of the permuted coach effects always are smaller than the original estimates. For example, our original estimate of coach effectiveness heterogeneity from the random-effects, lagged-score model and the external evaluator sample is 0.27 SD, whereas the 95th percentile of the permuted coach effects from the same sample and model is 0.19 SD. In the full sample, the 95th percentile of the permuted coach effects are no higher than 0.13 SD and substantially smaller than our original estimates (0.36 SD from gain-score model and 0.24 SD from the lagged-score model). Because all of our estimated coach effects are larger than the 95th percentile of coach "effects" across permutations, we reject the null that the true SD is zero.

Other Outcomes

Finally, we estimate coach effectiveness heterogeneity at improving two additional measures of teaching practice captured on TNTP's observation rubric. We do not include these measures in our primary analyses due to more severe ceiling effects compared to our primary measure. While Koedel and Betts (2010) find that severe right censoring has little practical implication for the estimation of teacher value added to student test scores in a set of simulations, we cannot say with certainty that these patterns extend to our context and estimation of coach effectiveness heterogeneity. As such, we cannot disentangle whether any differences in findings between measures is due to the underlying teaching constructs that the measures aim to capture or to measurement artifacts. Further, ceiling effects raise substantive concerns for school systems, coaches, and teachers about what exactly is being captured by the value-added estimates.

The two additional measures are: (i) *Culture of Learning*, which 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 (mean scores = 2.25 and 2.64 out of 3, for baseline and

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

outcome, respectively, in the preferred external evaluator sample; skew = −0.44 and −1.52, respectively; see online appendix table 4); and (ii) *Essential Content*, which 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 (mean scores = 2.34 and 2.71, for baseline and outcome, respectively, in the preferred external evaluator sample; skew = −0.62 and −1.68, respectively).

Relative to our preferred estimates of coach effectiveness heterogeneity for *Demonstration of Learning* (roughly 0.24 to 0.31 SD), estimates from our random-effects specification in the external evaluator sample are larger for *Essential Content* (0.28 to 0.41 SD) and smaller for *Culture of Learning* (0.13 to 0.22 SD). In some instances, estimates for the two additional outcomes are more sensitive to gain-score versus lagged-score specifications, compared with *Demonstration of Learning*. At the same time, overall patterns and sensitivity checks are similar. Inclusion versus exclusion of a lagged score has very little effect on results, even though there is more between-coach variation in baseline scores on *Culture of Learning* (ICC = 0.1 for baseline score and 0.04 for top-scoring teachers) and *Essential Content* (ICC = 0.06 for baseline score and 0.10 for top-scoring teachers) relative to *Demonstration of Learning* (ICC = 0.02 for baseline score and 0.03 for top-scoring teachers). Further, our estimates of coach effectiveness heterogeneity are not explained away by school fixed effects, nor by crossed rater-by-coach random effects. Additionally, permutation tests reject the null hypothesis of zero coach effects, with just one exception. For *Culture of Learning* in the external evaluator sample, the 95th percentile of the distribution of permuted effects estimated in a lagged-score specification and when randomization is conducted within site and across school years is slightly larger than our main estimate (0.17 versus 0.14 SD). For other permutation tests for *Culture of Learning* and all permutations for *Essential Content*, the 95th percentile is below our estimated coach effect.

7. 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 with the average effect of coaching programs. Using our preferred external evaluator sample, random-effects specifications, and outcome measure, we find that a 1 SD increase in coach effectiveness results in a roughly 0.25 to 0.3 SD increase in teacher performance, from gain-score and lagged-score models. A 2 SD increase in coach effectiveness results in a roughly 0.5 to 0.6 SD increase in teachers' instructional quality. Comparatively, meta-analytic estimates indicate that scaled-up instructional coaching programs improves teacher practice by 0.34 SD, which in turn translates into improvements in student test scores of roughly 0.1 SD (Kraft, Blazar, and Hogan 2018).

On one hand, our study's focus on teachers and coaches working in school districts across the United States increases generalizability relative to other similar studies conducted with a small number of coaches or in a single setting (Blazar and Kraft 2015,

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

2019). On the other hand, we focus only on the preservice 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 preservice teacher coaching has less coverage in the empirical literature base compared with in-service programs, recent experimental evidence of preservice 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).

There are a couple of limitations to our study that should be addressed in future research. To confirm our findings, additional studies might estimate coach effects under experimental conditions where coaches are randomly assigned to teachers. Our nonexperimental analyses identify some degree of nonrandom sorting on observables, though this does not appear to drive our main findings. At the same time, we cannot rule out the possibility of sorting on unobservables. Future studies may also probe in more depth assumptions related to statistical noise that we are not able to given features of our dataset. In particular, most coaches are observed for just one school year, meaning that there are relatively few teachers per coach. Model-based estimates of coach effectiveness heterogeneity that employ shrinkage partly address this concern. With more years of data per coach and more teachers per coach, it would be possible to apply a shrinkage factor that considers year-to-year variation in coach effects, which is an important consideration in the teacher value-added literature (T. J. Kane, Rockoff, and Staiger 2008; Bitler et al. 2021). The teacher value-added literature further suggests that two years of effectiveness estimates may only be weakly correlated, particularly with few students per teacher (McCaffrey et al. 2009), but that year-to-career correlations (making use of many years of data) increase substantially in magnitude. Thus, future studies may look for data where coaches are observed for at least three years. Notably, it would be difficult to pair this design with an experiment, as randomly assigning coaches to teachers over multiple years is less feasible than conducting random assignment in a single year. Future studies also should consider using an outcome measure that is less prone to censoring and ceiling effects.

To the extent that the magnitude of coach effectiveness heterogeneity is replicated in other studies that address limitations related to bias and noise, then there are directions for research that can inform policy and practice as well. For example, do estimates of coach effectiveness heterogeneity extend to the outcomes of students, who are the intended beneficiary of instructional improvement efforts? 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 is quite large and suggests that these relationships may further translate into changes in student outcomes. Future research may also examine coach effectiveness heterogeneity among a sample of individuals with more years of coaching experience than those at TNTP. It may be that coach effectiveness heterogeneity decreases with increased experience, as coaches improve in their work. Or, it could be that degrees of heterogeneity remain similar or grow in magnitude, as coaches build their own systems and structures for one-on-one work with teachers over time.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Another line 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? 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 and 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 that 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 (modeling, providing direct feedback, video observation, etc.) and technologies (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, McCombs, and Martorell 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. 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.

8. 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). (All cost estimates are adjusted to 2023 dollars.) However, these dollars generally are found to have very little, if any, return on investment (Yoon et al. 2007; Harris and Sass 2011; Fryer 2017). Coaching provides an attractive alternative, achieving some of the largest impacts on teacher and student

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

outcomes across all of the education intervention literature (Kraft, Blazar, and Hogan 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 to $10,500 per teacher per year. Examining coaching in an alternative-route teacher certification context, Kaufman et al. (2020) estimate that coaching constitutes 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 preservice 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 and 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 and training skilled coaches likely is doable and sustainable (Kraft, Blazar, and Hogan 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. Within TNTP's summer training context, the vast majority of coaches are hired as seasonal workers with limited coaching experience within the organization, likely due to the vast scale of the work. Other studies of scaled-up, statewide coaching programs also tend to hire and work with relatively novice coaches (Marsh, McCombs, and Martorell 2012). The inherent tradeoff between personnel quantity and quality also 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 (Stecher et al. 2001; Jepsen and Rivkin 2002). 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 whom 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. As 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

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

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, Blazar, and Hogan 2018), but will require strategic planning that focuses primarily on building a corps of highly skilled coaches.

ACKNOWLEDGMENTS

We thank our partners and collaborators at TNTP, including Vicky Brady and Bailey Cato, for compiling the data used in this project, and for ongoing brainstorming regarding analyses. We also thank Matthew Kraft, journal editors, and anonymous reviewers for providing valuable feedback on earlier drafts of the manuscript. The research reported in this article was supported by a grant from the Overdeck Family Foundation to TNTP. The opinions expressed are those of the authors and do not represent the views of the Foundation.

REFERENCES

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics* 3(3): 29–54. 10.1257/app.3.3.29

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Doug Staiger. 2019. An experimental evaluation of three teacher quality measures: value-added, classroom observations, and student surveys. *Economics of Education Review* 73(C). Available https://econpapers.repec.org/article/eeeecoedu/v_3a73_3ay_3a2019_3ai_3ac_3as0272775719302717.htm.

Bastian, Kevin, Kristina Patterson, and Dale Carpenter. 2020. Placed for success: Which teachers benefit from high-quality student teaching placements? *Educational Policy* 36(August): 0895904820951l2. 10.1177/0895904820951126

Battauz, Michela, Ruggero Bellio, and Enrico Gori. 2011. Covariate measurement error adjustment for multilevel models with application to educational data. *Journal of Educational and Behavioral Statistics* 36(3): 283–306. 10.3102/10769986l0366262

Bell, Courtney, Drew Gitomer, Daniel Mccaffrey, Bridget Hamre, Robert Pianta, and Yi Qi. 2012. An argument approach to observation protocol validity. *Educational Assessment* 17(April): 62–87. 10.1080/10627197.2012.715014

Bitler, Marianne, Sean P. Corcoran, Thurston Domina, and Emily K. Penner. 2021. Teacher effects on student achievement and height: A cautionary tale. *Journal of Research on Educational Effectiveness* 14(4): 900–924. 10.1080/19345747.2021.1917025

Blazar, David. 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. 10.1162/edfp_a_00251

Blazar, David, and Matthew A. Kraft. 2015. Exploring mechanisms of effective teacher coaching: A tale of two cohorts from a randomized experiment. *Educational Evaluation and Policy Analysis* 37(4): 542–66. 10.3102/0162373715579487

Blazar, David, and Matthew A. Kraft. 2019. Balancing rigor, replication, and relevance: A case for multiple-cohort, longitudinal experiments. *AERA Open* 5(3): 2332858419876252. 10.1177/2332858419876252

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Britton, Linda R., and Kenneth A. Anderson. 2010. Peer coaching and pre-service teachers: Examining an underutilised concept. *Teaching and Teacher Education: An International Journal of Research and Studies* 26(2): 306–314. 10.1016/j.tate.2009.03.008

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review* 26(6): 673–682. 10.1016/j.econedurev.2007.10.002

Cohen, Julie, Vivian Wong, Anandita Krishnamachari, and Rebekah Berlin. 2020. Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis* 42(2): 208–231. 10.3102/0162373720906217

Connor, Carol McDonald. 2017. Commentary on the special issue on instructional coaching models: Common elements of effective coaching models. *Theory Into Practice* 56(1): 78–83. 10.1080/00405841.2016.1274575

Darling-Hammond, Linda. 2017. Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education* 40(3): 291–309. 10.1080/02619768.2017.1315399

Denton, Carolyn A., and Jan Hasbrouck. 2009. A description of instructional coaching and its relationship to consultation. *Journal of Educational and Psychological Consultation* 19(2): 150–175. 10.1080/10474410802463296

Domina, Thurston, Ryan Lewis, Priyanka Agarwal, and Paul Hanselman. 2015. Professional sense-makers: Instructional specialists in contemporary schooling. *Educational Researcher* 44(6): 359–364. 10.3102/0013189X15601644

Education Next. 2018. EdStat: $18 billion a year is spent on professional development for U.S. teachers. *Education Next* (blog). 12 June 2018. Available http://www.educationnext.org/edstat-18-billion-year-spent-professional-development-u-s-teachers/.

Espinoza, Daniel, Ryan Saunders, Tara Kini, and Linda Darling-Hammond. 2018. Taking the long view: State efforts to solve teacher shortages by strengthening the profession. Learning Policy Institute. Available https://learningpolicyinstitute.org/product/long-view-report.

Ferrão, Maria Eugénia, and Harvey Goldstein. 2009. Adjusting for measurement error in the value added model: Evidence from Portugal. *Quality & Quantity* 43(6): 951–963. 10.1007/s11135-008-9171-1

Foote, Mary, Andrew Brantlinger, Hanna Haydar, Beverly Smith, and Lidia Gonzalez. 2011. Are we supporting teacher success: Insights from an alternative route mathematics teacher certification program for urban public schools. *Education and Urban Society - EDUC URBAN SOC* 43(May): 396–425. 10.1177/0013124510380420

Fryer Jr., Roland G. 2017. Management and student achievement: Evidence from a randomized field experiment. NBER Working Paper No. 23437. 10.3386/w23437

Goldhaber, Dan, John Krieg, and Roddy Theobald. 2019. Effective like me? Does having a more productive mentor improve the productivity of mentees? *Labour Economics* 63(December): 101792. 10.1016/j.labeco.2019.101792

Goldstein, Harvey, Daphne Kounali, and Anthony Robinson. 2008. Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling* 8(3): 243–261. 10.1177/1471082X0800800302

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Grissom, Jason A., Demetra Kalogrides, and Susanna Loeb. 2015. Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis* 37(1): 3–28. 10.3102/0162373714523831

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge. 2015. Can value-added measures of teacher performance be trusted? *Education Finance and Policy* 10(1): 117–156. 10.1162/EDFP_a_00153

Hanushek, Eric, and Steven Rivkin. 2010. Generalizations about using value-added measures of teacher quality. *American Economic Review* 100(2): 267–271. 10.1257/aer.100.2.267

Harris, Douglas N., and Tim R. Sass. 2011. Teacher training, teacher quality and student achievement. *Journal of Public Economics* 95(7–8): 798–812. 10.1016/j.jpubeco.2010.11.009

Harrison, Xavier A., Lynda Donaldson, Maria Eugenia Correa-Cano, Julian Evans, David N. Fisher, Cecily E. D. Goodwin, Beth S. Robinson, David J. Hodgson, and Richard Inger. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* 6(May): e4794. 10.7717/peerj.4794

Harville, David A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72(358): 320–338. 10.2307/2286796

Hill, Carolyn, Howard Bloom, Alison Black, and Mark Lipsey. 2008. Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives* 2(December): 172–177. 10.1111/j.1750-8606.2008.00061.x

Hill, Heather C., David Blazar, and Kathleen Lynch. 2015. Resources for teaching: Examining personal and institutional predictors of high-quality instruction. *AERA Open* 1(4): 2332858415617703. 10.1177/2332858415617703

Hill, Heather C., Charalambos Y. Charalambous, David Blazar, Daniel McGinn, Matthew A. Kraft, Mary Beisiegel, Andrea Humez, Erica Litke, and Kathleen Lynch. 2012. Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment* 17(2–3): 88–106. 10.1080/10627197.2012.715019

Holmstrom, Bengt, and Paul Milgrom. 1991. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7(special issue): 24–52. 10.1093/jleo/7.special_issue.24

Howley, Aimee Anton, Marged Howley Dudek, Rebekah Rittenberg, and William Larson. 2014. The development of a valid and reliable instrument for measuring instructional coaching skills. *Professional Development in Education* 40(5): 779–801. 10.1080/19415257.2014.919342

Jepsen, Christopher, and Steven Rivkin. 2002. *Class size reduction, teacher quality, and academic achievement in California public elementary schools.* San Francisco: Public Policy of Institute of California.

Joyce, Bruce, and Beverly Showers. 1981. Transfer of training: The contribution of "coaching." *Journal of Education* 163(2): 163–172. 10.1177/002205748116300208

Kane, Michael T. 2017. Measurement error and bias in value-added models: Measurement error and bias in VAMs. *ETS Research Report Series* 2017(1): 1–12. 10.1002/ets2.12153

Kane, Thomas, and Douglas Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. w14607. 10.3386/w14607

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review* 27(6): 615–631. 10.1016/j.econedurev.2007.05.005

Kane, Thomas J., and Douglas O. Staiger. 2012. Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*. Bill & Melinda Gates Foundation. https://eric.ed.gov/?id=ED540960

Kaufman, Julia H., Benjamin K. Master, Alice Huguet, Paul Youngmin Yoo, Susannah Faxon-Mills, David Schulker, and Geoffrey E. Grimm. 2020. *Growing teachers from within: Implementation, impact, and cost of an alternative teacher preparation program in three urban school districts*. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA256-1.html

Knight, David S., and Thomas M. Skrtic. 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. 10.1177/1052684620972048

Koedel, Cory, and Julian Betts. 2010. Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy* 5(1): 54–81. 10.1162/edfp.2009.5.1.5104

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. Value-added modeling: A review. *Economics of Education Review* 47(August): 180–195. 10.1016/j.econedurev.2015.01.006

Kraft, Matthew A. 2019. Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources* 54(1): 1–36. 10.3368/jhr.54.1.0916.8265R3

Kraft, Matthew A., David Blazar, and Dylan Hogan. 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. 10.3102/0034654318759268

Lockwood, J. R., and Daniel F. McCaffrey. 2014. Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics* 39(1): 22–52. 10.3102/1076998613509405

Marsh, Julie A., Jennifer Sloan McCombs, and Francisco Martorell. 2012. Reading coach quality: Findings from Florida middle schools. *Literacy Research and Instruction* 51(1): 1–26. 10.1080/19388071.2010.518662

Matsko, Kavita Kapadia, Matthew Ronfeldt, Hillary Greene Nolan, Joshua Klugman, Michelle Reininger, and Stacey L. Brockman. 2020. Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education* 71(1): 41–62. 10.1177/0022487118791992

McCaffrey, Daniel F., Tim Sass, J. R. Lockwood, and Kata Mihaly. 2009. The intertemporal variability of teacher effect estimates. *Education Finance and Policy* 4(4): 572–606. 10.1162/edfp.2009.4.4.572

McEachin, Andrew, Jonathan Schweig, Rachel Perera, and Isaac Opper. 2018. *Validation study of the TNTP core teaching rubric*. RAND Corporation. 10.7249/RR2623

Menzes, Ana, and Adam Maier. 2014. *Fast start: Training better teachers faster, with focus, practice and feedback*. TNTP. https://eric.ed.gov/?id=ED559704

Miles, Karen Hawley, Allan Odden, Mark Fermanich, and Sarah Archibald. 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.

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

Mulhern, Christine. 2022. Beyond teachers: Estimating individual guidance counselors' effects on educational attainment. EdWorkingPapers.Com. Annenberg Institute at Brown University. Available https://www.edworkingpapers.com/ai22-632.

National Center for Education Statistics (NCES). 2016. National teacher and principal survey (NTPS). Available https://nces.ed.gov/surveys/ntps/tables/Table_5_042617_fl_school.asp.

National Center for Education Statistics (NCES). 2020. Characteristics of public school teachers. May. https://nces.ed.gov/programs/coe/indicator_clr.asp.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. How large are teacher effects? *Educational Evaluation and Policy Analysis* 26(3): 237–257. 10.3102/01623737026003237

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods*, 2nd ed. Advanced Quantitative Techniques in the Social Sciences 1. Thousand Oaks, CA: Sage Publications.

Ronfeldt, Matthew, Emanuele Bardelli, Matthew Truwit, Hannah Mullman, Kevin Schaaf, and Julie C. Baker. 2020. Improving preservice teachers' feelings of preparedness to teach through recruitment of instructionally effective and experienced cooperating teachers: A randomized experiment. *Educational Evaluation and Policy Analysis* 42(4): 551–575. 10.3102/0162373720954183

Ronfeldt, Matthew, Stacey L. Brockman, and Shanyce L. Campbell. 2018. Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher* 47(7): 405–418. 10.3102/0013189X18782906

Russell, Jennifer Lin, Richard Correnti, Mary Kay Stein, Ally Thomas, Victoria Bill, and Laurie Speranzo. 2020. Mathematics coaching for conceptual understanding: Promising evidence regarding the Tennessee math coaching model. *Educational Evaluation and Policy Analysis* 42(3): 439–466. 10.3102/0162373720940699

Safran, D. G., D. A. Taira, W. H. Rogers, M. Kosinski, J. E. Ware, and A. R. Tarlov. 1998. Linking primary care performance to outcomes of care. *Journal of Family Practice* 47(3): 213–220.

Schochet, Peter Z., and Hanley S. Chiang. 2013. What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics* 38(2): 142–171. 10.3102/1076998611432174

Shannon, Darbianne K., Patricia A. Snyder, Mary Louise Hemmeter, and Mary McLean. 2021. Exploring coach–teacher interactions within a practice-based coaching partnership. *Topics in Early Childhood Special Education* 40(4): 229–240. 10.1177/0271121420910799

Shen, Jianping. 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. 10.2307/1164466

Slavin, Robert, and Dewi Smith. 2009. The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis* 31(4): 500–506. 10.3102/0162373709352369

Soland, James, Sara E. Rimm-Kaufman, Megan Kuhfeld, and Nadia Ventura-Abbas. 2022. Empirical benchmarks for changes in social and emotional skills over time. *Child Development* 93(4): 1129–1144. 10.1111/cdev.13741

Stecher, Brian, George Bohrnstedt, Michael Kirst, Joan McRobbie, and Trish Williams. 2001. Class-size reduction in California: A story of hope, promise, and unintended consequences. *Phi Delta Kappan* 82(May): 670–674. 10.1177/003172170108200907

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025

TNTP. 2014. *TNTP core teaching rubric: A tool for conducting common core-aligned classroom observations*. Available https://tntp.org/publications/view/tntp-core-teaching-rubric-a-tool-for-conducting-classroom-observations.

TNTP. 2018. *The opportunity myth: Technical appendix*. Available https://files.eric.ed.gov/fulltext/ED590222.pdf.

Walsh, Kate, and Sandi Jacobs. 2007. Alternative certification isn't alternative. Washington, DC: Thomas B. Fordham Institute. https://www.nctq.org/nctq/images/Alternative_Certification_Isnt_Alternative.pdf

Wenner, Julianne A., and Todd Campbell. 2017. The theoretical and empirical basis of teacher leadership: A review of the literature. *Review of Educational Research* 87(1): 134–171. 10.3102/0034654316653478

Wilson, Suzanne M. 2014. Innovation and the evolving system of U.S. teacher preparation. *Theory Into Practice* 53(3): 183–195. 10.1080/00405841.2014.916569

Wong, Kenneth K., and Anna Nicotera. 2006. Peer coaching as a strategy to build instructional capacity in low performing schools. In *Systemwide efforts to improve student achievement*, edited by Kenneth K. Wong and Stacey A. Rutledge, pp. 271–303. Research in Educational Policy : Local, National and Global Perspectives. Greenwich, CT: IAP - Information Age Pub.

Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley. 2007. *Reviewing the evidence on how teacher professional development affects student achievement. Issues & answers. REL 2007-No. 033*. Regional Educational Laboratory Southwest. Available https://eric.ed.gov/?id=ED498548.

Yopp, David A., Elizabeth A. Burroughs, John T. Sutton, and Mark C. Greenwood. 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. 10.1007/s10857-017-9373-3

Downloaded from http://direct.mit.edu/edfp/article-pdf/19/3/492/2456974/edfp_a_00407.pdf by UNIVERSITY OF MARYLAND COLLEGE PARK user on 27 February 2025