TN EDUCATION RESEARCH ALLIANCE

Cultivating Coaching in Clinical Mentors

An Experimental Evaluation of the Mentors Matter Professional Development Initiative

Matthew Ronfeldt, Matthew Truwit, Emanuele Bardelli, Kevin Schaaf, & Brian Smith

Abstract

Despite their critical role in the preparation of pre-service teachers, very little research has explored the impacts of offering mentors professional development around how to coach and support their candidates. We conduct an evaluation of a professional development opportunity offered by the state of Tennessee to randomly assigned mentors at six different programs over the course of three years, investigating its impacts on the perceptions and practices of both the mentors themselves and their candidates. We find that professional development increases mentors' frequency of coaching and mentors' own instructional effectiveness, particularly in emphasized areas, as well as candidates' employment rates, highlighting the potential of mentor professional development to improve the clinical placement experience as a whole.

WORKING PAPER
2022-01

This is a working paper. Working papers are preliminary versions meant for discussion purposes only in order to contribute to ongoing conversations about research and practice. Working papers have not undergone external peer review.

Acknowledgements: We appreciate the generous financial support that was provided for this research by the Institute of Education Sciences (IES), U.S. Department of Education through the Statewide, Longitudinal Data Systems Grant (PR/Award R372A150015). Emanuele Bardelli and Matthew Truwit also received pre-doctoral support from the Institute of Education Sciences (IES), U.S. Department of Education (PR/Award R305B150012). We are grateful to Kevin Schaaf for his work with the implementation and management of the Mentors Matter Recruitment initiative. This research was supported in part through computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor. This project would not have been possible without the partnership, support, and data provided by the Tennessee Department of Education and partner educator preparation providers. Any errors should be attributed to the authors.

Toward the end of their preservice preparation, aspiring teachers typically have a capstone clinical experience, often referred to as student teaching. For this experience, teacher preparation programs (henceforth programs) assign aspiring teachers, who we call teacher candidates, at least one clinical placement where they learn to teach P-12 students under the supervision of the teacher of record, who serves as a clinical mentor. Mentors usually help candidates develop as teachers not only by modeling effective instructional techniques but also by conducting formal observations and providing verbal and written feedback as candidates attempt to put into practice the knowledge and skills they have acquired through prior coursework and other preparation experiences.

Given that they often provide candidates their first real, hands-on opportunity to teach P-12 students in a classroom, clinical placements are considered particularly foundational experiences in the development of new teachers. Moreover, a wide and well-established body of research has illuminated the particular importance of the clinical mentor to candidate development (e.g., Goldhaber et al., 2020; Ronfeldt et al., 2018; Ronfeldt et al., 2020).[1]

Despite consensus around their significance, mentors report receiving very little professional development for how to most effectively foster candidates' development as teachers (Matsko et al., 2021). Consequently, one possible way to improve the clinical placement experience—for both mentors themselves and thereby their candidates—could involve the provision of professional development to mentors specifically focused on how to coach and support their candidates as they learn to teach.

Inspired by this possibility, the Tennessee Department of Education (TDOE) developed a series of professional development (PD) opportunities in early 2018 designed to cultivate high-

[1] See Ronfeldt (2021) for a full review of this literature.

quality coaching techniques in mentors, particularly with regard to the instructional practices of questioning and academic feedback. PD sessions aimed to provide mentors the chance to discuss and cultivate these two practices in order to both better support their candidates during their clinical placement and refine their own pedagogical techniques for the classroom. The TDOE then offered this PD over three successive academic years to a randomly assigned set of clinical mentors within a total of six different programs across the state.

Using both survey and administrative data, we are able to compare the outcomes of mentors assigned to receive PD with those of mentors in a business-as-usual condition, as well as the outcomes of the candidates coached by each group of mentors. We find that the provision of PD significantly influenced the coaching practices of mentors, increasing the frequency of those practices targeted specifically in the PD. Additionally, treatment increased the instructional effectiveness of mentors with their own P-12 students, concentrated again in the emphasized areas. Finally, we observe that the effects of the PD appear to extend beyond mentors, as candidates whose mentors were assigned to treatment found employment after graduation at significantly higher rates than their peers in control despite the fact that candidates in both experimental conditions reported feeling similarly prepared, satisfied, and inclined to teach at the end of their clinical placements.

Although encouraging mentors to take on additional responsibility by attending professional development posed some challenges that likely constrained the efficacy of this initiative, the results of this evaluation suggest that the use of PD to foster effective coaching practices has significant promise for improving the clinical placement experience for both mentors and candidates. Given the increasingly well-documented importance of high-quality clinical placements and mentorship to candidates, the adoption of thoughtfully designed PD

opportunities for teachers who wish to serve as mentors is a logical step in enhancing the development of the future teaching workforce.

Literature Review and Background

A number of descriptive correlational studies have collectively found that the quality of coaching provided by mentors predicts both candidates' self-perceived readiness to teach and their observed teaching effectiveness. Matsko et al. (2020), for example, administered surveys to all candidates and mentors in Chicago about their clinical experiences in general and specifically about the kinds, amount, and quality of coaching they felt they received and provided respectively. They found that candidates reported feeling better prepared to teach when they also reported that their mentors provided more frequent feedback, instructional support, collaborative coaching, and job-search support as well as a balance of autonomy and encouragement. However, candidates' feelings of preparedness were unrelated to their mentors' perceptions of the kinds, amount, and quality of coaching they provided.

Ronfeldt et al. (2021) extended this Chicago study by linking both candidates' and mentors' survey-based data on coaching to new outcomes: (1) mentors' perceptions of how well prepared to teach their candidates were at the end of their clinical experiences and (2) the observation ratings (of instructional performance) that recently-graduated candidates received from their principals during their first year of teaching in Chicago area schools. Consistent with the prior study, candidates who reported receiving more frequent and better-quality coaching from their mentors felt—and were rated by their mentors as—better prepared to teach; however, they had no better or worse first-year observation ratings after entering the classroom. When the authors considered the *mentors'* perceptions of the amount and quality of coaching they felt they offered to candidates, those candidates whose mentors reported providing more and better-

quality coaching in specific domains of instruction were rated as being better prepared to teach by their mentors and received better first-year observation ratings.

Given that mentors' coaching appears to have a positive relationship with candidates' perceived and actual readiness to teach, one might assume that teacher preparation programs would emphasize supporting and improving the coaching provided by those mentors working with their candidates. To the contrary, professional development focused on mentor coaching is relatively uncommon. For example, in a recent study extending the Chicago research described above, Matsko et al. (2021) found that only ten percent of mentors affiliated with traditional route programs in the Chicago area reported receiving professional development focused on how to coach candidates.[2]

Only two prior studies, to our knowledge, have examined the impacts of providing coaching professional development for mentors of preservice candidates (Becker et al., 2019; Giebelhaus & Bowman, 2002). Both studies found the professional development to positively impact both mentors' coaching practices and, in turn, candidates' teaching practices. Giebelhaus and Bowman (2002) randomly assigned 28 mentors from two programs to either professional development based upon principles and practices from the *Praxis III/Pathwise* framework or business-as-usual supports. Candidates paired with mentors who had been randomly assigned to the coaching professional development received higher observation ratings from trained, external evaluators than their peers in the control condition on most criteria. Becker et al. (2019) randomly assigned 130 mentors to one of three coaching professional development groups or a control group. They found that mentors who participated in professional development aligned their coaching practices to the training they received, according to their candidates. In addition,

[2] Rates were higher for mentors affiliated with alternative route (65 percent) and residency (58 percent) programs.

the candidates who worked with trained mentors not only reported receiving better quality coaching but were also evaluated as being more successful at addressing disruptive behaviors in their classrooms. However, substantial noncompliance issues post-randomization raise some concern about whether these effects can be interpreted causally.

Another study examined the impacts of coaching professional development for instructional coaches of teachers in an alternative teacher certification program (McQueen, 2013). Though not technically preservice candidates, the mentees were first-year teachers of record while still completing their certification requirements. McQueen (2013) randomly assigned half of all instructional coaches to receive professional development on how to provide choice-based, focused coaching around a single targeted instructional practice. Compared to mentees whose coaches were assigned to the business-as-usual condition, those with coaches assigned to the professional development reported better quality coaching and earned better observation ratings.

The prior literature then suggests that coaching professional development can positively impact mentors' coaching and candidates' teaching. However, these studies also stand to benefit from extension in a number of directions, each of which the present research addresses. First, most prior work focused on a relatively small number of coaches, candidates, and/or programs raising questions about whether the same coaching professional development would be viable at scale. Relatedly, these studies were generally underpowered. Our study examines the effects of coaching professional development offered to almost 500 mentors across six programs. Additionally, the coaching professional development we study differs from that offered in prior research in that it goes beyond fostering mentors' coaching practices to also target their teaching practices, given our hypothesis that—by improving their teaching practices—mentors will be

better equipped to coach their candidates, especially in the skill areas that are the focus of professional development. As a result, our research also goes beyond prior studies in testing whether professional development impacts the teaching practices of mentors themselves and not just those of the candidates they coach. Finally, our study extends prior work to consider candidates' employment as an outcome measure, a possible additional signal for candidates' readiness to teach.

The Mentors Matter Professional Development Initiative

The TDOE set out to create a PD specifically aligned to high-quality coaching practices for mentors. Drawing on much of the research literature on coaching and teacher preparation cited above, the PD was developed under the leadership of the department's educator evaluation coaches and its divisions of educator effectiveness, preparation, and research, all in partnership with staff from the Nashville Teacher Residency.

Together, these stakeholders identified two specific objectives for the PD—encouraging mentors to both model questioning practices for their candidates and provide them with related actionable feedback rooted in data from observations. These predominant emphases stemmed from the historically low performance of teachers across the state—especially early career teachers—in the pedagogical practice of questioning.[3] Sessions included a balance of didactic instruction and active practice. Moreover, the PD provided opportunities for mentors not only to foster the development of effective questioning and feedback practices in *candidates*' direct interactions with students but also to rehearse effectively using these same skills *themselves* both when observing and debriefing their candidates and in the classroom with their own students.

[3] Specifically, all educators tend to score lower on the state rubric's *questioning* indicator relative to other indicators. On top of its significant room for growth, *questioning* is also considered by officials at the TDOE to be a high-leverage instructional practice.

In the spring of 2018, the TDOE began offering the PD to mentors, piloting the sessions with two programs. The following academic year, the department partnered with three programs (including one from the prior year) to implement the PD at a larger scale, a process they repeated again in the 2019-20 academic year with two of the largest programs in the state.[4]

In its pilot semester, the treatment consisted of two face-to-face sessions, each of which took place over the course of a full workday on site at one of the partner programs. The department's educator evaluation coaches led the delivery of the first day of PD. Through modeling, discussing, and practicing effective questioning techniques, the first day of the PD aimed to strengthen mentors' skills in accurately evaluating their candidates' questioning, particularly in terms of gathering evidence around the strength of candidate practices, in order to craft next steps to help candidates improve. The second day of PD was led by the executive director of the Nashville Teacher Residency and focused on improving feedback practices. The second day of PD featured opportunities for mentors to engage in role-playing in order to practice and refine how to provide feedback to candidates. The learning also emphasized the importance of mentors targeting their improvement efforts by identifying one big takeaway from each observation as well as the value of providing candidates with multiple opportunities to practice their skills and receive feedback.

Each year of implementation also included the further development and revision of the PD based on feedback from participating mentors and partner programs, as shown in Figure 1. Across implementation years, feedback from mentors attending the PD was generally very positive; participants reported having a high-quality PD experience that they would highly recommend to their peers (see Descriptive Appendix for analyses of post-session questionnaires).

[4] We refer to these three implementations as the "Pilot Semester", "Year 1", and "Year 2", respectively.

However, in response to the need for more support in structuring their observation procedures professed by mentors during the pilot semester, the team of PD developers created a suite of guides designed to facilitate the planning, gathering of evidence during, and reflecting on observations of candidate practice. These guides served to provide mentors with a structure for (1) leading planning conversations with their teacher candidate, (2) observing lessons and gathering evidence, and (3) conducting reflective post-observation conversations with their teacher candidate and providing feedback. Guides were revised in Year 2, including a streamlining of the PD for better alignment, with the aim being to ensure that mentors completed the PD equipped to effectively implement the strategies encapsulated within the guides independently.

Over the same period of time, the PD also evolved significantly to supplement its two in-person sessions with online webinars and asynchronous practice assignments. Prior to the first face-to-face gathering, mentors participated in an introductory webinar (either live or via recording) in which the initiative, the handbook, and the guides were all introduced. Additionally, after practicing during the in-person sessions, mentors were asked to complete and upload multiple sets of the planning, evidence-gathering, and reflection guides over the course of the semester, using Canvas, an online learning management platform. A concluding webinar was also added to wrap up the PD at the end of the clinical placement experience.

Research Questions

As evaluators of this initiative, we sought to answer the following questions:

1. Did the PD directly impact the coaching and teaching practices of mentors?
2. Did the PD indirectly impact the self-perceptions and workforce outcomes of candidates?

Methods

Sample

The initiative involved 380 candidates and 474 mentors from six different partner programs over the course of three academic years. Since participation in the initiative did not alter programs' existing clinical placement practices, candidates had either one or two mentors depending on their programs' placement policies. Similarly, placement durations ranged from under eight weeks to the entire academic year and had varying degrees of commitment and responsibility.

To account for these program differences, we randomly assigned candidate-mentor dyads within each program to either a treatment (i.e., inviting their mentor(s) to attend PD) or control condition. We evaluate the effectiveness of our randomization by checking for balance between candidates and mentors in treatment and control conditions on a set of observable characteristics in Table 1. Our joint chi-squared test finds no evidence of significant difference between the two groups overall (X^2 (37, N = 380) = 27.16, p = .88), although the treatment group appears to comprise a smaller proportion of candidates preparing to teach at the secondary level. We consequently control for placement school grade level in all analytic models.

After randomization, 14.0% of candidate-mentor dyads dropped out over the course of the study. Attrition from the initiative was either due to candidates' individual circumstances (e.g., Praxis failure, changing majors, deferring student teaching for personal reasons, accepting a full-time teaching job mid-year, etc.) or by school districts opting out of the initiative altogether after randomization (e.g., district leaders no longer felt feasible for some of their teachers to attend the training during the school day). While dyads who dropped out were marginally more likely to have been assigned to treatment (6.3 percentage points, p = .08), the combination of differential and overall attrition is deemed tolerable under both optimistic and

cautious assumptions (What Works Clearinghouse, 2020). Moreover, nearly all attrition occurred in the period between randomization and the actual administration of treatment, making it unlikely that participation in treatment induced participants to drop out and suggesting that differential attrition behavior between conditions does not pose a threat to the internal validity of our experiment. Moreover, a parallel balance check by treatment condition of the 327 candidates and 396 mentors in the post-attrition sample finds nearly identical results to that using the full sample after randomization, suggesting little impact of participant characteristics on attrition.

We also examine the differences between treatment dyads by actual PD attendance, comparing attending and absent mentors and their candidates on the same observable characteristics. As shown in Table 2 below, our omnibus test finds evidence of significant differences between the two groups overall at baseline. Specifically, mentors who attended PD had higher likelihoods of having already met and worked with their candidates prior to pre-survey administration, higher ORs in the *professionalism* domain, and fewer years of experience; in addition, their candidates felt more prepared at baseline, possibly as a result of already having a chance to begin working with their mentors.[5] Participation rates also varied substantially by program and year. We unpack the potential implications of these differences when discussing compliance and treatment effects in the Discussion section below.

Data

Our data come from two main sets of sources—a pair of surveys administered to all candidates and three statewide administrative datasets from Tennessee's state longitudinal data

[5] Though pre-surveys were sent early in the fall before candidates officially began their clinical placements, our partner programs typically matched candidates to mentors and placements in the spring. In most cases, mentors and candidates had already made contact over the summer and some had even already begun collaborating.

system—as well as a small supplementary set of audio recordings of some treatment mentors' coaching conversations, each of which we elaborate on below.

Survey Data

We surveyed all candidates at the beginning (pre-survey) and end (post-survey) of their clinical experiences.[6] On both pre- and post-surveys, candidates were asked about their feelings of preparedness to engage in a wide variety of teaching skills as well as their intended career plans for the upcoming year. Both surveys also included a series of questions about candidates' demographics (e.g., race, gender, age, etc.) and academic history (e.g., GPA, program, prior educational experience, etc.). Candidates were also asked on the post-survey about the frequency of and their satisfaction with the coaching and feedback they received from their mentors. The pre-survey additionally included items asking if candidates had previously met and/or worked with their mentors. Items on the pre-survey were incorporated into balance checks and as controls, when appropriate; those from the post-survey were created into outcome measures, which we discuss later.

63.6% of all candidates ($N = 208$) responded to the pre-survey, 50.8% ($N = 166$) to the post-survey, and 40.4% ($N = 132$) to both. These middling response rates stem mostly from substantially lower survey completion in the second year of the initiative, likely in part due to the onset of the COVID-19 pandemic. Candidate response patterns to the post-survey, pre-survey, and both surveys did not differ significantly by treatment condition. Additionally, the sample of dyads whose candidates responded to the post-survey and on which we focus for many of our analyses remained both balanced overall by treatment condition and similar to those with non-responding candidates across the observable characteristics (see Appendix Table 1).

[6] We surveyed candidates with more than one placement at the conclusion of their final placement.

It is worth acknowledging that we surveyed mentors as a part of this initiative as well. Mentors, who were surveyed once at random over the course of the clinical placement, were similarly asked to report on the frequency of coaching practices used with and feedback provided to their candidates in the preceding week (or a typical week if they had not coached in the prior week). Mentors were also asked sets of questions about their experiences serving as mentors, and their likelihood to serve again in the future. Finally, as with candidates, we also inquired about mentors' demographic information and backgrounds.

The mentor survey response rate (54.3%, $N = 215$) was comparable to that for candidates (despite mentors solely receiving compensation for participation). However, mentor response rates differed more substantially by treatment condition. Across all three years of the initiative, mentors assigned to treatment were 10.8 percentage points ($p = .02$) more likely to complete a survey than those assigned to control, with those who attended any PD 28.5 percentage points ($p < .001$) more likely than those who did not. Furthermore, a joint chi-squared test comparing the sample of mentors who completed a survey with those who did not found marginally significant overall differences ($X^2(23, N = 283) = 31.96$, $p = .10$), largely due to the higher ORs and VAMs of responding mentors. As a result, given concerns about external and internal validity due to the unexceptional overall response rate and the differential response pattern by treatment condition and PD attendance, we opt to only use the demographic and background information from the mentor survey as covariates and eschew the other variables as outcomes.[7]

Administrative Data

[7] In the interest of transparency, we briefly present results from analysis of mentor survey outcomes here. We estimate the effects of the PD on mentors' self-reported (a) frequency of coaching provided to their candidates, (b) value gained from serving as a mentor, and (c) likelihood of serving again. No treatment effect on any of these perceptions reached a level of significance, although most trended slightly negative. We recommend interpreting these with substantial caution for the reasons mentioned above. We also discuss (a) in greater detail below.

We also make use of three different statewide administrative datasets provided by TDOE. First, we rely on a dataset used primarily for teacher compensation and retirement for elements of teachers' employment—including school(s), teaching assignment(s), and salary—as well as demographic information such as race/ethnicity, gender, age, years of experience, and highest level of educational attainment. These variables are included mostly in balance checks and as controls when appropriate.

We additionally incorporate two other data sources, each containing a different measure of teachers' instructional effectiveness: observation ratings (ORs) and value-added measures (VAMs).[8] In general, we use measures of instructional effectiveness from the years concurrent and subsequent to the initiative as outcomes, while we incorporate those in the year prior to the initiative as controls; however, the COVID-19 pandemic placed significant constraints on the availability of these workforce measures, which we touch on more later.

Across all years, we are able to match 89.9% of mentors ($N = 356$) with any of these three administrative datasets. These match rates for each individual variable range from as low as 28.5% for VAMs[9] to as high as 85.1% for ORs. Overall, there is no difference in the match rates across all variables for mentor administrative data by treatment condition (X^2 (12, $N = 396$) = 13.40, $p = .34$).

With regard to candidates, it is not possible to determine a true "match rate" given that we cannot distinguish those who end up working in Tennessee public schools but do not appear in the data from those who end up working in private schools, moving to a different state, or

[8] VAMs are designed to isolate and quantify an individual teacher's impact on student achievement through lagged growth models. Notably, Tennessee has a state-specific approach to calculating value-added that differs from other VAMs in that it does not directly account for students' demographic characteristics but instead is based solely on students' performance on state tests in prior years (SCORE, 2014; Vosters et al., 2018).

[9] This low match rate is expected given that only teachers in tested subjects (i.e., those teaching third through eighth grade and certain core high school subjects) receive VAMs.

failing to obtain employment.[10] However, we have no reason to expect that the true match rate should differ by randomly assigned treatment condition ex ante nor do we believe that treatment would differentially impact our capacity to match candidates (contingent on their being hired).

Audio Recording Data

Finally, in the most recent year of the initiative, treatment mentors who attended PD were asked to record and submit a brief post-observation conference held with their candidates. These conversations typically lasted less than 10 minutes. We incorporate the contents of 29 recorded coaching conversations as a secondary data source to briefly supplement our analysis of mentor coaching practices in the Results section below.

Outcome Measures

To evaluate the impacts of the PD, we construct one set of outcome measures based on candidates' self-reported post-survey responses and another set from statewide administrative data sources.

Survey

Using items from the post-survey, we developed eight standardized outcome measures through confirmatory factor analysis based on the survey's intended structure, as well as a ninth outcome in a simpler fashion (described below).[11] We provide a qualitative description of each factor included in our analyses below, with the psychometric properties of each reported in greater detail in the Technical Appendix.

Coaching Frequency. We measure the frequency of coaching candidates reported receiving across four sub-factors of coaching practices best described as common, data-driven, collaborative, and modeling. Common coaching practices include two items asking candidates

[10] In essence, we treat the match rate for candidates as an employment rate.
[11] The factor outcomes follow analogous structures to those first described in Ronfeldt et al. (2020).

about the frequency of their mentors' observations and prompts to practice specific aspects of teaching practice. Data-driven coaching practices draw on six items that focus on how often mentors used data from observations or student work to guide coaching. Collaborative coaching practices comprise two items that focus on co-planning and co-teaching activities. Modeling coaching practices contain two items that focus on how often a mentor modeled specific instructional strategies for the candidate. We also calculate an average of these four sub-factor scores to measure candidates' overall perceptions of the frequency of coaching provided by their mentors.

Satisfaction. We assess candidates' satisfaction with the mentoring they received across two sub-factors, one centered on the helpfulness of the coaching candidates received from their mentors and the other focused on the supportiveness of the environment their mentors cultivated.[12] The *helpful coaching* sub-factor includes nine items that assess candidates' satisfaction with the quality and frequency of the specific coaching practices and feedback of their mentors. The *supportive environment* sub-factor includes four items that measure the extent to which candidates felt comfortable asking their mentors for help or taking risks in front of their mentors. We similarly calculate an average of these two sub-factor scores to measure teacher candidates' overall satisfaction with the coaching provided by their mentors.

Preparedness. We measure candidates' perceptions of their own preparedness on both the pre- and post-survey. We divide this construct into two correlated sub-factors: preparedness in questioning skills – which were an explicit focus of the PD for mentors – and preparedness in other instructional skills. The first sub-factor includes five items assessing how prepared candidates felt to independently develop, plan, and ask questions to engage students in

[12] In previous work (e.g., Ronfeldt et al. (2020)), we referred to these subfactors as "support/feedback" and "autonomy/engagement" respectively.

understanding a concept. Given the emphasis of the PD on the practice of questioning, we deemed it important to examine preparedness in this area separately. The second sub-factor includes six items asking about other aspects of planning for and delivering instruction unrelated to questioning, such as developing materials, providing examples or analogies for new concepts, and incorporating multimedia into a lesson. We also calculate an average of these two sub-factor scores to measure candidates' overall feelings of preparedness for independent instruction.

Plans to Teach. Finally, candidates were also asked on both the pre- and post-survey about their career plans after graduation, including whether they planned to teach next year (as opposed to working in education in a capacity apart from teaching, working in a field other than education, not working at all, going to graduate school, or being uncertain about the next year). For simplicity, we created a binary variable that indicated whether a candidate planned to teach in the year after program completion or not and employed it as either an outcome (when created from the post-survey) or a control (when from the pre-survey).

Workforce Outcomes

We originally intended to examine the impacts of the PD across a much broader set of workforce outcomes for both mentors and candidates, including multiple measures of instructional effectiveness for each. However, these planned analyses were substantially complicated by the outbreak of COVID-19, which resulted in the cancelation of a year of state testing, a pause on observations of teachers for a little over a year, and the subsequent loss or compromising of OR and VAM data for many mentors and hired candidates.[13] As a consequence

[13] More specifically, we only have VAM outcomes for mentors in tested subjects from the first two years ($N = 65$, 16.4%), VAM outcomes for employed candidates in tested subjects from the pilot semester ($N = 14$, 4.3%) and OR outcomes for employed candidates from the first two years ($N = 91$, 27.9%). As a result, we choose to omit analyses using these measures as outcomes given our lack of confidence in their precision and validity, though in the interest of transparency, we still report the general findings of analysis of these outcomes here. The offer of PD appeared to have a slight negative impact on the VAMs of this subset of mentors equal to -0.32 student-level standard deviations

of this unanticipated and considerable reduction in sample size, we found ourselves limited to examining only a pair of workforce measures as outcomes—mentors' ORs and candidate employment.[14]

Observation Ratings (ORs). All teachers in Tennessee public schools are observed multiple times per year using the TEAM rubric, a tool for evaluating instructional effectiveness across 23 indicators that range from "Significantly Below Expectations" (1) to "Significantly Above Expectations" (5).[15] These indicators are contained within (and are averaged to calculate scores for) four (4) broader domains: *instruction*, *planning*, *environment*, and *professionalism*. The values for these domains are then averaged again to calculate an overall teacher OR for the year on a similar 1 to 5 scale.[16] We primarily examine mentors' year-average ORs at three different levels: (1) overall, (2) in the domain of *instruction*, and (3) on the specific indicator of *questioning*. We choose these latter two given that they capture the instructional practice(s) most emphasized in the PD sessions.

Employment. To measure candidates' employment rates, we construct a binary variable indicating whether candidates were eventually hired (or not) into a public school in the state of Tennessee. Any candidate who appears in the teacher compensation dataset or has at least one of the two evaluation metrics (ORs and/or VAMs) in any year subsequent to that of participation in the initiative is considered hired; all other candidates are classified as unhired. We acknowledge

(p = .05) but did not appear to influence the ORs of hired candidates. We again encourage caution when interpreting these estimates given their imprecision as well as their small and distinct subsample.

[14] We do include mentor VAMs from the year prior to their participation in the initiative as covariates in our balance checks given that these data were unaffected by the pandemic.

[15] While some districts use alternative approved evaluation rubrics, the state translates these to TEAM-equivalent scores for teacher evaluation purposes, which we use in our analyses.

[16] We construct overall year-average ORs in this manner so as to maintain logical consistency with the domain-level averages. However, an alternative option could involve simply averaging across all items regardless of domain to create an overall year-average OR. The correlation between overall ORs calculated using these two approaches is 0.98, with nearly identical results produced using each estimate.

that this operationalization may result in marking some candidates who successfully found employment after program completion as "unhired" (e.g., those missing data, those in private schools, those in other states, etc.). However, as state law requires the evaluation of all teachers in Tennessee's public schools, the TDOE is confident that these datasets identify the vast majority of public school teachers employed in the state. Furthermore, given randomization, we assume that the likelihood of incorrect classification of teachers as (un-)employed is comparable across treatment conditions, allowing us to therefore attribute any differences in match rates by treatment condition to real differences in employment caused by treatment. Consequently, while our estimates of the absolute level of candidates' successfully finding employment after graduation may be too low, our estimates of the difference in employment rates by treatment status should be accurate and internally valid.

Analysis

The within-program random assignment of candidate-mentor dyads to each treatment condition allows for a relatively simple analytic approach to estimating the effect of *offering* the PD. However, given a mentor participation rate of roughly 50%, we also find it valuable to provide a complementary estimate of the impact of actually *attending* any PD on each outcome.

Intent-to-Treat (ITT)

In our ITT model, we calculate the effect of being assigned to treatment for candidates and mentors. These estimates do not distinguish between compliant and noncompliant mentors (and/or their candidates) but instead provide an estimate of the impact of simply inviting mentors to the PD. We calculate our ITT estimates using Equation 1 shown below.

$$Y_{ij} = \beta_0 + \beta_1 \cdot Treat_{ij} + \Delta_i \cdot B + \gamma_j + \epsilon_{ij} \quad (1)$$

Y_{ij} is a given outcome for candidate or mentor i at program x cohort j, γ_j is a cohort (i.e., program x year) fixed effect, Δ_i is a vector of candidate-level controls that address imbalances in treatment conditions for placement level (i.e., elementary or secondary) and alternative certification pathway,[17] and ϵ_{ij} is a heteroskedastic-robust error term. Additionally, when exploring the effects of the PD specifically on the outcome of mentors' ORs, we include an additional control for the instructional effectiveness of mentor i at program x cohort j in the year prior to the initiative.

Given the random assignment of candidate-mentor dyads, the minimal threat of attrition, and the balance on baseline characteristics previously reported, this model produces credibly causal estimates of the PD's effects. At the same time, the ITT may be an overly conservative estimate for understanding the value of the PD since it only captures the effect of mentors being *invited* to attend. Given that only about half of the invited mentors participated, the ITT may therefore understate the PD's actual impact on the mentors who *attended* and prioritize internally valid estimates over the perhaps more externally valid estimates produced by a model that accounts for mentor compliance. For this reason, we also calculate local average treatment effect (LATE) estimates that adjust our ITT estimates by the fraction of participants that complied with their experimental assignment and attended PD.

Local Average Treatment Effect (LATE)

[17] While this latter variable is not significantly imbalanced at randomization, we believe it is important to include for two reasons: (1) its obvious relevance to employment as an outcome, given that alternatively certified candidates are typically already employed during their student teaching placements, and (2) the theoretically relevant difference in the coaching provided by mentors of job-embedded candidates as opposed to their traditional peers.

The LATE offers a relevant estimate of the effect of mentors actually attending the PD.[18] Calculating the LATE requires the use of a two-stage least squares instrumental variables (IV) approach, as shown in Equation 2 below.

$$Attend_{ij} = \beta_0 + \beta_1 \cdot Treat_{ij} + \Delta_i \cdot B + \gamma_j + \epsilon_{ij} \quad (2)$$

$$Y_{ij} = \alpha_0 + \alpha_1 \cdot \widehat{Attend} + \Delta_i \cdot A + \delta_j + \phi_{ij} \quad (3)$$

In the first stage (Equation 2), we use the random assignment of treatment to predict each mentor's likelihood of attending any PD. Then, in the second stage (Equation 3), we incorporate these estimated probabilities as our primary predictor of interest in calculating the effect of actually attending any PD. In essence, by using our random assignment as an instrument for mentor participation, we end up scaling our ITT estimates by the inverse of the mentor participation rate.

An IV approach, however, typically requires significantly stricter assumptions than the simpler ITT in order to produce valid causal estimates of a treatment effect. Specifically, these include (1) *relevance*, or an association between instrument (random assignment) and treatment (attending PD); (2) *exclusion*, or no direct pathway between instrument and outcomes except through treatment; (3) *exchangeability*, or no common causes or confounders shared between instrument and outcomes; and (4) *monotonicity*, or the empirically untestable assumption that likelihood of receiving treatment is not higher for any individual assigned to control than to treatment.

However, given the random assignment in our experimental design, these assumptions are easily satisfied either empirically or intuitively. *Relevance* is satisfied given that the

[18] For models involving candidate post-survey or workforce outcomes, we consider candidates with two placements as having had a mentor attend PD if either mentor was recorded as participating in any face-to-face session.

invitation to attend PD was only offered to individuals in treatment, and since being invited to—but not attending PD—would be highly unlikely to impact any of the outcomes of interest, the *exclusion* restriction is also met. Our use of random assignment and the results of our initial balance check in Table 1 both provide strong evidence of *exchangeability* and ensure that the distribution of hypothetical compliers would be unlikely to differ across treatment conditions. As a result, we present results from the LATE models as illustration of both the magnitude of the impacts of actually attending the PD and the greater potential of the PD to influence practice and perceptions were more mentors induced to participate.

Results

We begin by examining whether the offer of and attendance at the PD directly influenced mentor practices, examining first candidates' perceptions of the coaching they received and then mentors' own classroom instruction. We then turn to whether the PD indirectly impacted candidates' perceptions of their own preparedness and intentions to teach as well as their likelihood of finding employment after graduation.

Mentor Coaching Practices

The first panel of Table 3 displays both the intent-to-treat (ITT) and local average treatment effect (LATE) estimates for the respective effects of offering and attending PD on the coaching that mentors provided according to their candidates. We find that candidates in treatment (i.e., those with mentors invited to attend the PD) reported receiving a third of a standard deviation more frequent coaching overall than their peers in control. This effect was driven largely by an increased frequency of the coaching practices emphasized specifically during the PD sessions, including data-driven practices (0.44 standard deviations, $p = .007$), like asking reflective questions and sharing specific next steps for improvement, and common

practices (0.43 standard deviations, $p = .009$) such as conducting observations and prompting specific candidate practices. When exploring the effect of actually attending the PD, these estimates are essentially doubled (as, intuitively, they are scaled up by the inverse of the participation rate of roughly 50%). Candidates whose mentors attended at least one face-to-face component of the PD reported receiving between 0.66 ($p = .03$) and 0.87 ($p = .007$) standard deviations more frequent coaching along the same dimensions than their peers whose mentors did not participate.[19]

While evidence of the PD's impact on the *quantity* of coaching that candidates reported receiving is substantial, the PD's influence on the perceived *quality* of mentors' coaching is less clear. Candidates in treatment appear to have felt slightly more satisfied (roughly a fifth of a standard deviation) with the coaching they received from their mentors, both in terms of the quality of support given and the degree of autonomy granted by their mentors. However, this result did not reach a level of statistical significance.

Qualitative Analysis of PD Uptake in Coaching Conversations

In pursuit of confirmatory evidence of the high prevalence of training-encouraged coaching practices reported by treatment candidates, we conducted an ancillary qualitative analysis of coaching conversations recorded and submitted by 29 trained mentors in Year 2 (17.5% of all Year 2 dyads, 35.0% of those assigned to the treatment condition, and 59.2% of

[19] We also asked questions about the frequency of coaching on the aforementioned mentor survey. Interestingly, we found no evidence of an effect of the PD on the frequency of coaching that mentors self-reported providing their candidates. However, we are skeptical of the results that draw on this mentor survey for a number of reasons: (1) significantly differential response rates by treatment condition and PD participation as previously described; (2) prior research that suggests that candidates' responses typically provide greater insight into mentors' actual coaching practices, as mentors uniformly tend to score themselves higher across dimensions of coaching (Matsko et al., 2021); and (3) empirical support for this trend, with mentor coaching frequency measures both higher on average and with less variation than candidate coaching frequency measures, regardless of treatment condition or compliance status.

those who attended the PD).[20] We transcribed and analyzed these coaching conversations using a combination of a priori and open coding; a team of coders based the a priori codes on the goals of the PD as identified through a review of materials and resources provided to PD participants while also developing a set of emergent codes that fell outside the context of the PD's goals (see Appendix Table 2). A primary and secondary coder then applied final codes to each coding unit[21] in a given transcript with acceptable levels of agreement (pooled kappa = 0.83).

This qualitative analysis of recorded coaching conversations from treated candidate-mentor dyads offers some opportunity to investigate the actual (and not just reported) coaching practices used by mentors. However, as they are limited to a small and non-representative sample that lacks a comparison group, these analyses do not intend to make claims about the effects of treatment on coaching frequency. Instead, they intend to provide a description of the coaching practices used by mentors who attended PD while coaching their candidates and explore whether and how these practices reflected elements of the PD.

Progression of Coaching Conversations. We began by working to develop an understanding of the typical nature and structure of mentors' coaching conversations with candidates. We found that the typical progression of most coaching conversations closely resembled the guides that the PD provided to mentors, which emphasized recalling specific classroom instances, connecting them to candidates' practice, and then determining next steps.

[20] It is worth noting that this subsample differs significantly from their peers who did attend PD but did not submit recordings. Dyads in the audio recording subsample were more likely to come from the elementary level; mentors were more likely to be older, White, and female, with higher average levels of experience and observation ratings; finally, candidates were more likely to be higher-achieving, experienced, female, and undergraduate as well as more likely to have had contact with their mentors at the start of their placements. As such, we suggest caution in viewing these coaching conversations as representative of all participating dyads and in interpreting findings as an influence of PD.

[21] Coding units were first separated by turns of talk and then by focal topic. For example, if one speaker changed topic within their turn of talk, this turn of talk was separated into two coding units.

Most mentors began the substantive portions of the conversations by asking their candidates to recall specific instances from their lessons when they noticed that students were succeeding or struggling with the day's objectives. Following candidates' responses, mentors tended to provide additional instances themselves or ask candidates to provide further evidence of their own. Mentors would then typically shift to inquiring about which specific moves candidates had made that might have led to the particular success or struggle of the students. This line of questioning not only grounded the conversation in the specific events of the lessons but also prompted candidates to connect student actions with their own. In the cases where candidates and their students struggled, mentors would then usually move on to questions that focused on determining next steps to improve the practices of their candidates.

Frequency and Enactment of Coaching Practices. We continued our investigation by calculating the frequencies with which all codes appeared in the data to evaluate which dimensions of coaching were most prevalent throughout these conversations (see Appendix Table 3). We looked at the specific conversational moves that candidates and mentors made as well as the focal content of the conversations.

Of particular note here are the frequencies for the *Probing Questions* (conversational move) and *Questioning* (focal content) codes. The *Probing Questions* code was applied to any coding unit containing a question aimed to elicit a response from the candidate. Nearly 45% of coding units in which the mentor is speaking (as well as 24.5% of all coding units) included a probing question. This suggests that mentors worked to meaningfully engage candidates in the process of debriefing and reflecting upon their practice in observed lessons, the predominant focus of the PD. Meanwhile, the *Questioning* code was applied whenever the mentors and candidates discussed the instructional practice of questioning students, which we observed in

26.3% of all coding units. This relatively high prevalence also aligns with the specific content of the PD, given its strong emphasis on improving and modeling the questioning practices of candidates.

Another explicit focus of the PD was to encourage mentors to provide actionable feedback to candidates. Two codes (*Feedback - Perspective* and *Feedback - Data Driven*) were designed to capture the particular types of feedback provided by mentors. Perspective feedback was subjective (e.g., "I thought your lesson went well…") and occasionally provided insight from the mentor (e.g., "That comes from additional practice and experience"), whereas data-driven feedback contained quantitative (e.g., "I noticed you called on 3 boys and 5 girls") or qualitative (e.g., "I noticed you used a lot of turn and talk") data drawn from observation of the lesson. Our analyses indicate that mentors provided both types of feedback to their candidates frequently, though they were twice as likely to provide feedback that offered perspective (31.3%) as opposed to feedback that directly referenced data collected during the lesson (15.7%).

Finally, from this feedback, mentors and candidates worked together to establish next steps for the practice of the candidate. The *Recommendation* code was applied in instances where either speaker suggested a practice or strategy for the candidate to utilize in the future, which occurred in roughly 17% of all coding units. Further analyses of the *Recommendation* code revealed that mentors (46%) and candidates (54%) each made roughly half of these suggestions, indicating that mentors and candidates consistently constructed next steps for candidates' practice in collaboration. This collaborative practice coincides with the guides provided during PD that focused the attention of post-observation conversations on co-constructing goals and planning next steps for candidates' future practice.

Together, these analyses suggest a substantial degree of fidelity in takeup of the PD among mentors. The conversations show a strong content focus on instructional practices—questioning, in particular—over other aspects of teaching practice like unit planning, classroom management, or professional responsibilities, which we might have expected to see highlighted otherwise, given that these are often a typical concern among PSTs and first-year teachers (Matsko et al., 2020; Veenman, 1984). In addition, mentors' use and progression of particular conversational strategies to engage candidates may reflect that the emphases and resources of the PD were embraced in the coaching repertoires of participating mentors. Mentors consistently asked probing questions to candidates, which they then followed with feedback grounded in data from the lesson, closely resembling the guides from the PD. These findings may offer supporting evidence of the ways in which the foci and tools highlighted in the PD have the capacity to influence mentors' coaching practice. However, since we did not analyze recordings of control mentors, we cannot conclude whether this resemblance was a product of the training or simply a universal structure for coaching conversations across all mentors.

Mentor Teaching Practices

The second panel of Table 3 summarizes the results of our exploration of whether the PD also impacted mentors' instructional practice in the classroom with their own students, given that sessions aimed to improve mentors' own teaching (of P12 students) in addition to their coaching (of candidates). We look at the PD's effects on mentors' year-average observation ratings (ORs) at three levels - overall, in the domain of *instruction*, and on the specific indicator of *questioning* - in both the years concurrent and subsequent to our initiative.[22] For both, we see modest but

[22] While the concurrent year's ORs are closer in proximity to the PD, some or all of the individual observations that comprise this annual average may have occurred before the sessions themselves. Using the subsequent year's ORs avoids this issue but relies on observations of instructional practice that may have been quite distant from the sessions and also results in the sample loss of mentors from the third year of implementation because of data

imprecisely estimated improvements in the instructional effectiveness of mentors' questioning of their own students after being invited to the PD; we also see a similar, consistent pattern for mentors' overall and *instruction* ORs in the subsequent year. These estimates in Panel B of Table 3 arguably reach practical, if not statistical, significance; for example, the 0.12 point increase in mentors' *questioning* ORs in the year subsequent to the PD is equivalent to one sixth of a standard deviation and comes at a point in most mentors' careers (i.e., more than 10 years on the job) when ORs typically plateau or grow only incrementally. In fact, these gains are roughly equivalent to the difference in principals' evaluations of rapidly developing first- and second-year teachers and much larger than the typical minimal returns to experience expected of teachers after ten or more years teaching (Kraft, Papay, & Chi, 2020). Furthermore, when examining the effects of actually attending the PD, these estimates are quite large (e.g., ranging from a third of a standard deviation in overall OR to half of a standard deviation in *questioning*) though they still do not reach statistical significance.[23]

We run a number of different checks to help ascertain the robustness of these imprecisely estimated effects of the PD on mentors' instructional effectiveness. First, we conduct a placebo test in which we use the same modeling specification to calculate treatment effects for the other 22 indicators on the state observation rubric that were less emphasized by the PD and then rank these effects from largest to smallest. For concurrent year ORs, the treatment effect for *questioning* is the largest of all 23 indicators; in the subsequent year, *questioning* is the sixth largest. Separately, we also create a new dataset at the level of the date of each individual

limitations due to COVID-19. Estimates of the effects of the PD on concurrent year ORs for the subsample of mentors for whom we have subsequent year data (i.e., those in the first two years of implementation) are somewhat smaller than those for the whole sample, suggesting that, if anything, the reduction in sample size contributes a negative bias to our estimate of the effect on subsequent year ORs for all mentors.

[23] The models that produce these estimates control for mentors' overall, *instruction*, and *questioning* ORs in the year prior to the PD; estimates are reduced slightly without such adjustments.

observation (rather than examining mentors' year-average ORs). We then use a generalized difference-in-differences approach to estimate the increase in ORs between the pre- and post-PD periods for mentors in treatment above and beyond the baseline increase over the same time span for their peers in control. These alternative estimates of the impact of being invited to the PD on ORs for both *instruction* ($\beta = 0.06$) and *questioning* ($\beta = 0.10$) are slightly smaller than those for the subsequent year ITT in Table 4 but are marginally significant ($p < .10$ for both).[24] A comparable placebo test using this specification reveals that the estimate for *questioning* is the third largest of all the indicators.[25] Finally, we estimate event study models as a complement to these difference-in-difference specifications and find no evidence of any dynamic treatment effect in the period leading up to the PD (i.e., pre-trend) on either *instruction* or *questioning*.[26] Altogether, the consistency of findings across this suite of robustness checks suggests that the observed effects of treatment on mentors' instructional effectiveness - especially in the *instruction* domain and on the *questioning* indicator - is not purely due to chance.

Candidate Self-Perceptions of Preparedness and Career Plans

Given that the PD appeared to have statistically significant and sizable effects on the frequency of mentors' coaching of candidates, as well as marginally and practically significant impacts on their instructional practices, we then began to explore whether it might also have indirect effects on the perceptions and performance of their candidates. We address our second

[24] We also employ a triple difference approach that further interacts both our post-PD indicator and treatment status with actual attendance at PD to obtain a treatment on the treated (TOT) estimate as a rough comparison for the LATE estimates in Table 3. This model produces estimates of the effect of actually attending the PD on *instruction* ($\beta = 0.24$) and *questioning* ($\beta = 0.34$) that are again slightly reduced but still quite substantial in magnitude and marginally significant (this time for *instruction* only).

[25] *Questioning* also has the largest treatment effect of all indicators in the placebo test for the above triple difference specification.

[26] We specify time as the running variable for these models in two different ways - both ordinally in relation to time of the first session (e.g., first, second, and third observations before and after) and continuously by the number of calendar days between the observation and the first session. Our conclusions are qualitatively similar regardless of specification.

research question by presenting estimates of the impact of being invited to and attending the PD on candidate outcomes in Table 4. Overall, as shown in the first panel, the PD did not appear to have any significant effects on candidates' perceptions of their experiences at the conclusion of their clinical placements. That is, compared to peers in control, candidates in treatment reported statistically similar levels of preparedness to independently engage in a variety of instructional skills and intentions of entering teaching after graduation, even when accounting for those same perceptions and intentions prior to the beginning of their clinical placements.

Candidate Workforce Outcomes

As mentioned above, our plans to move beyond candidates' perceptions and examine the PD's effects on their own instructional effectiveness (i.e., ORs and value-added) once on the job were unfortunately hampered by the COVID-19 pandemic. Because of the pause in student testing and observation of teachers, candidates' rates of employment are the sole workforce outcome for which data are available.

In the second panel of Table 4, we find that the offer of PD to mentors resulted in a 14.5 percentage point increase in candidates' likelihood of ever finding a public school teaching position in the state of Tennessee ($p = .005$). Moreover, candidates who had mentors that actually attended any face-to-face sessions were 28.9 percentage points more likely to find employment at some point after program completion ($p = .005$).[27] As a follow-up, we restricted our analysis to only those candidates who indicated on their pre-survey that they planned to seek employment as a teacher after graduation (as opposed to other jobs in education, graduate school, etc.). Effects of the PD on employment were twice as large ($\beta_{ITT} = 0.26$, $p < 0.001$; $\beta_{LATE} = 0.59$, $p < 0.001$).[28]

[27] We find slightly smaller but still statistically significant effects of the PD on candidate employment if restricting the outcome to finding employment in the year immediately after program completion.

[28] About half of this increase, however, stems from first restricting the analysis to the subsample of candidates who responded to the pre-survey at all ($\beta_{ITT} = 0.21$, $p = 0.001$; $\beta_{LATE} = 0.44$, $p = 0.001$).

It is not initially clear how offering PD to mentors could result in candidates' increased likelihood of employment. We can think of two possible categories of explanations – that the PD altered *mentors'* behavior supporting candidates in the job-seeking process (e.g., by increasing mentors' investment in and support of candidates' pursuit of employment) or that it impacted *candidates*' performance in ways that increased their attractiveness on the job market (e.g., with the increase in coaching improving candidates' practice, resulting in more positive letters of recommendation from principals or more effective performance during job interviews).

One way we can begin to distinguish between these two broad possibilities is to explore *where* treatment candidates most often successfully found employment. For instance, if treatment candidates were disproportionately hired into the same schools as their mentors, we might lend more credence to the possibility that treatment induced mentors to more strongly advocate on behalf of hiring their candidates (and question whether the PD actually increased candidates' attractiveness on the job market); on the other hand, if treatment candidates were more successful at finding jobs in entirely different districts from their mentors than their control peers, it would be harder to believe that mentors were directly responsible for this effect (and more convincing that it somehow stemmed from the candidates themselves).

Therefore, we conduct further exploratory analyses to determine whether candidates in treatment found employment in the same schools as, same districts (but not schools) as, or entirely distinct settings from their mentors. Using a multinomial logistic regression with each of these three settings as mutually exclusive outcomes, we find that treatment candidates' increased likelihood of employment overall is driven by increased rates of hiring both in mentors' schools and in different districts altogether (but not at a statistically significant level in different schools in mentors' districts). Given that the majority (56.7%) of candidates who find employment in

either treatment condition do so in different districts than their mentors, the results of these analyses suggest that this candidate employment treatment effect is unlikely to be fully explained by changes in mentors' support of and advocacy for candidates on the job market.

Discussion

Given the considerable evidence illustrating the influence of high-quality mentors (Goldhaber et al., 2020; Ronfeldt et al., 2018; Ronfeldt et al., 2020), the development of an effective PD designed to improve the coaching practices of mentors is an obvious potential lever for enhancing the preparation of early career teachers. However, since it is not typically offered to mentors (Matsko et al., 2020), very few studies have explored if PD focused on coaching actually produces any tangible benefit. While the existing research suggests that offering PD to mentors can improve the coaching practices of mentors and the instructional effectiveness of their candidates, it has been small in scale and in highly unique settings; it is unclear if such PD will have similar impacts across heterogeneous teacher education contexts. The results of this study therefore provide valuable insight into one state's scaled-up approach to improving the clinical preparation experiences across six programs, producing causal estimates of the effects of mentors being invited to and attending a PD designed to foster effective coaching practices on both mentor and candidate outcomes.

The most important takeaway we find is that—at least according to candidates—PD impacts the coaching practices of mentors. PD increased the frequency with which mentors engaged in almost all kinds of coaching; moreover, these improvements were largely concentrated in areas that were explicit focuses of the PD sessions, which were selected due to being both historically challenging for early career teachers and foundational instructional practices for candidates (e.g., asking reflective questions, providing specific feedback,

constructing next steps for improvement, etc.). The magnitudes of these effects – beginning at two thirds of a standard deviation and ranging higher – match the substantial estimates of PD's impact on mentor practices in prior literature (Becker et al., 2019; McQueen, 2018), especially when focusing on the mentors who actually attended any face-to-face component of the PD. There is also some suggestive evidence that candidates with mentors who were offered PD may have felt slightly more satisfied with the quality of coaching they received, though these effects are not statistically significant.

The PD also seems to have benefited mentors in terms of their own instructional effectiveness. Our findings suggest that mentors who were offered the PD, especially after accounting for their performance in the preceding year, became slightly more instructionally effective with regard to questioning their own students, the main instructional focus of the PD. This is also supported by a pattern observed in the open-ended responses of mentors on post-session questionnaires wherein many mentors cited the dual benefits of the PD—both to their candidates and to their own P12 students (see Descriptive Appendix for examples). Though effect sizes are modest, it is worth keeping in mind that the mentors in our sample have on average twelve years of teaching experience, a point at which growth in instructional effectiveness has typically reached a plateau; as such, we believe that these effects are both noteworthy and of practical significance.

However, we do not find any conclusive evidence that the PD, or the resultant increase in the frequency of mentor coaching, produces any impact on candidates' perceptions of their own preparedness or plans to teach in the next year. One possible explanation for why changes in the frequency and focus of mentors' coaching practices might not have translated into changes in candidates' feelings of preparedness and career plans may simply be that the effects of the PD

were limited to mentors and did not reach candidates. For example, the magnitude of the increased frequency in mentors' coaching practices may simply not have been enough to significantly impact candidates' feelings of readiness to teach or career plans. Alternatively, while the PD seemed to elicit *more* coaching, it may not have necessarily been more *effective* coaching, at least for inducing a candidate to feel more prepared or more likely to pursue teaching. In general, candidates' perceptions are likely influenced by many aspects of clinical mentors beyond their coaching—as well as many aspects of clinical preparation beyond mentors—all of which make it more challenging and less likely to observe any distal effects of the PD. However, counter to these explanations, we find suggestive evidence of a modest increase in candidate satisfaction with coaching, as well as strong evidence of an indirect impact of offering PD to mentors on candidates' likelihood of finding employment after graduation; moreover, this latter effect was most concentrated in different districts than their clinical placements, where mentors are least likely to directly contribute to the hiring of their candidates.

Therefore, another explanation might be that the PD did, in fact, positively impact candidates' performance and attractiveness on the job market through their mentors in some way, even if candidates reported feeling no more satisfied or better prepared. For example, lessons about questioning and feedback learned in PD by mentors and passed down to candidates may have improved treatment candidates' performance during interviews despite these candidates not feeling any more confident after or content with their clinical placement experience. Consistent with this explanation, some research has shown that candidates' self-reported preparedness has little relationship with their actual instructional effectiveness in the first year of teaching; in other words, candidates who are actually more prepared may not necessarily feel better prepared (Ronfeldt et al., 2020). More specifically, while the actual preparedness of candidates in

treatment may have increased, their standards for success or their anticipation of the challenges of teaching may have similarly risen, resulting in no relative difference in their self-evaluation of their own readiness when compared to less prepared but also less conscientious peers. Overall, understanding the complex connections between both the frequency and quality of mentor coaching and the many dimensions of candidate perceptions and performance could offer valuable insight for how and where to most effectively improve existing mentors' practices and their relationships with candidates; however, it is unfortunately beyond the capacity of this experiment – though a valuable direction for future study – to truly disentangle and quantify these complicated mechanisms.

While we find that the introduction of this PD opportunity largely appeared to benefit mentors and candidates, we acknowledge a number of limitations to this work. First, analyses involving candidate perceptions as outcomes rely on survey completion. Low overall response rates due to challenges in collecting data through partner programs, as well as some differential participation by treatment condition and compliance, present concerns about both the internal and external validity of survey results, though we believe that our checks of the representativeness and balance of the analytic sample mitigate most of these. Furthermore, it is worth emphasizing that our survey measures are self-reported perceptions. For example, as mentioned above, we do not present results on candidates' *actual* instructional effectiveness (aside from employment, which is at best a rough proxy) but rather their own *perceived* preparedness, largely due to the loss of OR and VAM data during the COVID-19 pandemic. Consequently, we strongly encourage future research that examines the impacts of PD for mentors on candidates' teaching performance and whether these impacts mediate or moderate candidates' success finding employment.

Moreover, we do not examine the effects of PD on mentors' actual coaching but rather their candidates' perceptions of the coaching they received. Our qualitative analysis of the audio recordings of mentors' coaching conversations provides initial evidence that observed coaching practices are largely consistent with the survey, but data were limited to a subsample of mentors who attended the PD in the final year that does not include any mentors from the control condition. Future research should, therefore, consider incorporating observational measures for the coaching practices of mentors as well—and ideally those of mentors in both treatment and business-as-usual conditions.

Above all, though, the primary challenge with the implementation of this study involved mentor attendance at the PD. Even in our most well-managed years of the initiative, we had difficulty inducing a majority of mentors to participate in the PD. One possibility is that the treatment mentors who decided to attend PD sessions differed in significant ways from those who did not, and that these differences might explain part of their decision to participate or not. In fact, the balance check presented in Table 2 that compares candidates and mentors assigned to treatment by actual PD attendance did find evidence of significant differences between attending and absent dyads at baseline. As a result, it is possible that the kinds of mentors who were not willing or able to attend the PD with its current structure and incentives may be more challenging to recruit and might even require a different strategy altogether to facilitate or encourage their participation. For example, these mentors, who appear older and may be less likely to take on additional leadership responsibilities, may already be stretched too thin with teaching and family responsibilities to commit to a time-intensive, in-person professional development.

Consequently, increasing PD attendance might require considerable adjustment to the structure or incentives of the PD. For example, transitioning the synchronous, face-to-face

sessions to a virtual setting may ease the burden on mentors already at capacity, a hypothesis supported by the higher attendance rate in Year 2 which did so during the pandemic (50% compared to 25% in prior implementations). Additionally, though the financial compensation for attending the face-to-face sessions was substantial and many logistical constraints were addressed by the department's offers of substitute coverage and travel reimbursement, it may be that these types of incentives simply do not address the reasons for mentors' low attendance rates. For example, PD sessions were scheduled during the work day, and while TDOE provided substitute teachers for mentors, some teachers simply might prefer not attending the PD over having a substitute teacher in the classroom. We encourage any future research involving the professional development of mentors to explore what structures and/or what types of incentives and supports are most appropriate and effective for the large responsibility of coaching the future teaching workforce shouldered by mentors.

Despite the challenges inherent in fostering effective coaching practices in mentors through in-person PD, we still believe this approach has substantial promise for improving the clinical experiences of candidates. As mentors often receive little to no guidance on how best to coach their candidates, PD focused on mentoring practices appears to be an area ripe for development and future research. Moreover, the magnitude and variety of effects estimated in this study illustrate the potential benefits to mentors themselves, their candidates, and all of their students combined, particularly if the structure of the PD can be optimized. Receiving more frequent and targeted coaching, especially around foundational instructional practices in particularly challenging domains, is in and of itself a tangible benefit for candidates without even considering that it appears to increase the likelihood that candidates' find employment. Moreover, such PD appears to additionally serve as an opportunity for mentors to hone their own

instructional practices for the benefit of their classroom students. As research continues to highlight the value of high-quality mentorship during candidates' clinical placements, the development of PD designed to cultivate coaching in mentors offers a clear avenue for improving the preparation experiences of the future teaching workforce.

References

Becker, E. S., Waldis, M., & Staub, F. C. (2019). Advancing student teachers' learning in the teaching practicum through Content-Focused Coaching: A field experiment. *Teaching and Teacher Education, 83*, 12-26. https://doi.org/10.1016/j.tate.2019.03.007

Giebelhaus, C. R. & Bowman, C. L. (2002). Teaching mentors: Is it worth the effort? *The Journal of Educational Research, 95*(4), 246-254. https://doi.org/10.1080/00220670209596597

Goldhaber, D., Krieg, J., & Theobald, R. (2020). Effective like me? Does having a more productive mentor improve the productivity of mentees? *Labour Economics, 63*, 101792. https://doi.org/10.1016/j.labeco.2019.101792

Kraft, M.A., Papay, J.P., & Chi O.L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, *39*(2), 315-347. https://doi.org/10.1002/pam.22193

Matsko, K.K., Ronfeldt, M., & Greene Nolan, H. (2021). How different are they? Comparing preparation offered by traditional, alternative, and residency pathways. *Journal of Teacher Education*. Online First: https://doi.org/10.1177/00224871211015976

Matsko, K. K., Ronfeldt, M., Greene Nolan, H., Klugman, J., Reininger, M., & Brockman, S. L. (2020). Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education, 71*(1), 41-62. https://doi.org/10.1177/0022487118791992

McQueen, K. (2018). *Promoting instructional improvement: Promising evidence of coaching that benefits teachers' practice* [Doctoral dissertation, University of Michigan]. Deep Blue.

Ronfeldt, M. (2021). *Links among teacher preparation, retention, and teaching effectiveness*. National Academy of Education. https://doi.org/10.31094/2021/3/1

Ronfeldt, M., Bardelli, E., Mullman, H., Truwit, M., Schaaf, K., & Baker, J. (2020). Improving student teachers' readiness to teach through recruitment of instructionally effective and experienced cooperating teachers: A randomized experiment. *Educational Evaluation and Policy Analysis, 42*(4), 551-575. https://doi.org/10.3102/0162373720954183

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher, 47*(7), 405-418. https://doi.org/10.3102/0013189X18782906

Ronfeldt, M., Matsko, K. K., Nolan, H. G., & Reininger, M. (2021). Three different measures of graduates' instructional effectiveness and the features of preservice preparation that predict them. *Journal of Teacher Education, 71*(1), 56-71. https://doi.org/10.1177/0022487120919753

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research, 54*(2), 143-178.

Table 1. Balance Check of Candidate-Mentor Dyad Characteristics by Treatment Condition

Full Sample	All	Control	Treatment	Diff.	Std. Diff.	*N*
Panel A: Candidate Characteristics						
Female	0.87	0.85	0.89	0.04	0.12	286
White	0.94	0.92	0.96	0.03	0.15	285
Undergraduate	0.75	0.75	0.75	0.01	0.01	239
GPA	3.60	3.63	3.58	-0.05	0.16	151
Age	24.1	24.3	24.0	-0.29	0.05	256
Alternative pathway	0.08	0.10	0.06	-0.04	0.13	380
Elementary	0.72	0.71	0.74	0.04	0.08	378
Secondary	0.41	0.46	0.36	-0.10	0.20^{+}	378
Prev. met mentor	0.87	0.88	0.86	-0.02	0.07	241
Prev. worked with mentor	0.11	0.08	0.14	0.05	0.17	208
Initial preparedness	-0.08	-0.12	-0.05	0.07	0.07	224
Questioning	-0.09	-0.13	-0.05	0.08	0.08	224
Other	-0.08	-0.10	-0.05	0.05	0.05	224
Plans to teach next year	0.87	0.87	0.86	-0.02	0.05	224
Has prior ed. experience	0.48	0.47	0.48	0.01	0.03	257
Panel B: Mentor Characteristics						
VAM (*overall*)	0.38	0.34	0.41	0.07	0.10	115
English	0.26	0.30	0.21	-0.09	0.11	60
Math	0.54	0.38	0.70	0.32	0.41	43
OR (*overall*)	4.31	4.30	4.31	0.01	0.03	305
Instruction	4.18	4.17	4.19	0.01	0.03	298
Environment	4.64	4.64	4.65	0.01	0.03	297
Planning	4.19	4.19	4.19	0.00	0.00	298
Professionalism	4.54	4.55	4.54	-0.01	0.03	306
Questioning	4.05	4.08	4.03	-0.05	0.08	298
Years of experience	12.9	12.8	13.0	0.14	0.02	307
Total salary	$48,022	$48,299	$47,718	-$580.65	0.06	311
Age	41.8	41.5	42.1	0.59	0.06	305
White	0.97	0.98	0.97	-0.02	0.09	306
Female	0.89	0.87	0.90	0.02	0.07	305
Master's degree or higher	0.62	0.65	0.58	-0.07	0.14	309
Previously mentored	0.82	0.79	0.84	0.05	0.12	228
Previously received PD	0.28	0.33	0.24	-0.09	0.19	198
Panel C: Program Characteristics						
Year 1	0.19	0.18	0.19	0.01	0.02	380
Year 2	0.30	0.31	0.30	-0.01	0.03	380
Year 3	0.51	0.51	0.51	0.01	0.01	380

Program 1	0.12	0.12	0.13	0.01	0.02	380
Program 2	0.08	0.08	0.08	0.00	0.02	380
Program 3	0.06	0.06	0.05	0.00	0.02	380
Program 4	0.23	0.23	0.23	0.00	0.01	380
Program 5	0.27	0.27	0.27	0.00	0.01	380
Program 6	0.24	0.24	0.24	0.00	0.01	380
Number of placements	1.25	1.25	1.24	-0.01	0.02	380
Yearlong placement	0.57	0.57	0.57	0.00	0.00	380

Note. Joint chi-squared test for 380 observations with 37 degrees of freedom = 27.16, $p = .88$. When candidates had more than one mentor, continuous mentor characteristics (e.g., salary) are calculated as averages and binary characteristics (e.g., white) as maxima (i.e., if-ever). Coverage/missingness patterns do not differ significantly by treatment condition. $^{+}$ $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 2. Balance Check of Treatment Dyad Characteristics by PD Attendance

Full Sample	All	Absent	Attending	Diff.	Std. Diff.	*N*
Panel A: Candidate Characteristics						
Female	0.88	0.87	0.89	0.02	0.05	134
White	0.96	0.98	0.94	-0.04	0.21	134
Undergraduate	0.75	0.78	0.74	-0.04	0.09	114
GPA	3.60	3.61	3.60	-0.01	0.03	77
Age	24.0	24.3	23.8	-0.48	0.09	125
Alternative pathway	0.07	0.09	0.05	-0.03	0.14	165
Elementary	0.78	0.80	0.77	-0.03	0.07	164
Secondary	0.34	0.35	0.33	-0.02	0.05	164
Prev. met mentor	0.86	0.81	0.90	0.09	0.25	118
Prev. worked with mentor	0.13	0.08	0.16	0.08	0.25	99
Initial preparedness	-0.06	-0.28	0.09	0.37	0.37^{+}	110
Questioning	-0.07	-0.30	0.09	0.39	0.40^{*}	110
Other	-0.06	-0.27	0.08	0.35	0.33^{+}	110
Plans to teach next year	0.86	0.89	0.85	-0.04	0.12	110
Has prior ed. experience	0.44	0.37	0.49	0.13	0.26	127
Panel B: Mentor Characteristics						
VAM (*overall*)	0.38	0.26	0.45	0.19	0.25	50
English	0.18	0.03	0.29	0.26	0.34	25
Math	0.75	1.02	0.65	-0.37	0.43	18
OR (*overall*)	4.32	4.31	4.32	0.01	0.03	137
Instruction	4.20	4.20	4.20	0.00	0.00	132
Environment	4.63	4.65	4.62	-0.02	0.06	132
Planning	4.18	4.21	4.17	-0.04	0.08	132
Professionalism	4.54	4.47	4.60	0.13	0.31^{+}	135
Questioning	4.02	4.02	4.01	-0.01	0.02	132
Years of experience	12.8	14.6	11.4	-3.16	0.36^{*}	137
Total salary	$47,719	$48,972	$46,772	-$2,199.2	0.20	137
Age	41.7	43.7	40.3	-3.32	0.33^{+}	135
White	0.96	0.98	0.95	-0.03	0.18	136
Female	0.89	0.89	0.88	-0.01	0.03	135
Master's degree or higher	0.57	0.60	0.55	-0.05	0.10	138
Previously mentored	0.83	0.85	0.82	-0.03	0.08	121
Previously received PD	0.24	0.24	0.24	0.01	0.02	112
Panel C: Program Characteristics						
Year 1	0.25	0.22	0.27	0.05	0.12	165
Year 2	0.25	0.29	0.22	-0.07	0.16	165
Year 3	0.50	0.49	0.51	0.02	0.04	165

Program 1	0.15	0.10	0.19	0.09	0.24	165
Program 2	0.12	0.14	0.09	-0.05	0.16	165
Program 3	0.06	0.04	0.07	0.03	0.12	165
Program 4	0.17	0.22	0.14	-0.08	0.22	165
Program 5	0.22	0.29	0.18	-0.11	0.27^{+}	165
Program 6	0.28	0.20	0.33	0.13	0.29^{+}	165
Number of placements	1.19	1.25	1.15	-0.10	0.26	165
Yearlong placement	0.56	0.54	0.58	0.05	0.09	165

Note. Joint chi-squared test for 165 observations with 37 degrees of freedom = 54.10, $p = .03$. When candidates had more than one mentor, continuous mentor characteristics (e.g., salary) are calculated as averages and binary characteristics (e.g., white) as maxima (i.e., if-ever). Coverage/missingness patterns do not differ significantly by treatment condition. $^{+}$ $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 3. Treatment effects on mentor practice

	Invited to PD (ITT)	Attended PD (LATE)
Panel A: Candidate Perceptions of Mentor Coaching		
Coaching Frequency	0.331*	0.660*
	(0.151)	(0.299)
Common	0.426**	0.848**
	(0.162)	(0.321)
Data-Driven	0.438**	0.872**
	(0.159)	(0.318)
Collaborative	0.113	0.225
	(0.166)	(0.330)
Modeling	0.349*	0.696*
	(0.162)	(0.322)
Coaching Satisfaction	0.187	0.364
	(0.163)	(0.319)
Support	0.184	0.358
	(0.164)	(0.321)
Autonomy	0.190	0.369
	(0.166)	(0.326)
N	166	166
Panel B: Observation Ratings (ORs)		
	Concurrent Year	
Overall	-0.003	-0.007
	(0.033)	(0.080)
Instruction	0.005	0.013
	(0.036)	(0.087)
Questioning	0.107	0.256
	(0.067)	(0.162)
N	319	319
	Subsequent Year	
Overall	0.063	0.194
	(0.051)	(0.164)
Instruction	0.089	0.282
	(0.055)	(0.182)
Questioning	0.120	0.382
	(0.096)	(0.315)
N	157	157

Note. All models include fixed effects for program *x* cohort and school level (e.g., elementary). Coaching frequency models also include a fixed effect for candidates in alternative certification pathways, while OR models include controls for ratings in the year

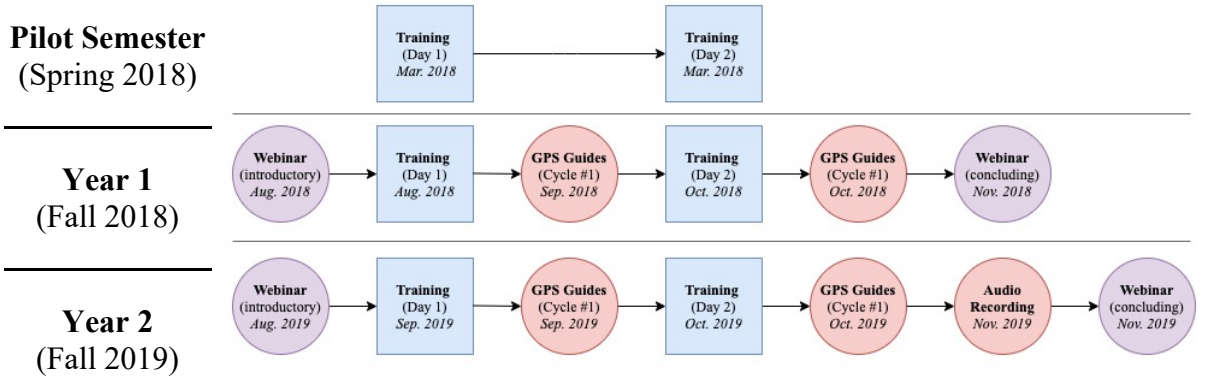
prior to the initiative. Coaching frequency measured in standard deviations; ORs on a 5-point scale. Robust standard errors in parentheses. $^{+}$ $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 4. Treatment effects on candidate outcomes

	Invited to PD (ITT)	Attended PD (LATE)
Panel A: Candidate Self-Perceptions		
Feelings of Preparedness	0.082	0.164
	(0.167)	(0.333)
Questioning	0.118	0.234
	(0.169)	(0.338)
Other	0.047	0.094
	(0.173)	(0.346)
Plans to Teach	-0.032	-0.064
	(0.040)	(0.081)
N	158	158
Panel B: Workforce Outcomes		
Employed	0.145**	0.289**
	(0.051)	(0.102)
N	326	326

Note. All models include fixed effects for program *x* cohort, school level (e.g., elementary), and alternative certification pathways. Satisfaction and preparedness measured in standard deviations; plans to teach and hired in percentage points. Robust standard errors in parentheses. $^{+}$ $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure 1. Evolution of PD Components by Implementation Year



Note. Square PD components attended in-person; circular completed virtually. Blue components offered synchronously, red asynchronously, and purple both.

Appendix

Appendix Table 1. Balance Check of Survey Responding Dyad Characteristics by Treatment

Candidate Post-Survey Sample	All	Control	Treatment	Diff.	Std. Diff.	*N*
Panel A: Candidate Characteristics						
Female	0.86	0.84	0.88	0.05	0.13	161
White	0.93	0.88	0.96	0.08	0.31^{+}	160
Undergraduate	0.78	0.77	0.80	0.03	0.08	125
GPA	3.69	3.59	3.58	-0.01	0.03	83
Age	24.6	24.6	24.6	0.01	0.00	158
Alternative pathway	0.09	0.10	0.08	-0.02	0.07	166
Elementary	0.73	0.65	0.80	0.16	0.36*	166
Secondary	0.42	0.52	0.32	-0.20	0.40**	166
Prev. met mentor	0.86	0.90	0.83	-0.07	0.19	137
Prev. worked with mentor	0.14	0.13	0.14	0.01	0.03	118
Initial preparedness	-0.08	-0.02	-0.13	-0.11	0.12	132
Questioning	-0.10	-0.04	-0.15	-0.11	0.12	132
Other	-0.06	0.00	-0.11	-0.11	0.11	132
Plans to teach next year	0.89	0.89	0.90	0.00	0.01	132
Has prior ed. experience	0.50	0.47	0.53	0.06	0.12	160
Panel B: Mentor Characteristics						
VAM (*overall*)	0.29	0.23	0.34	0.10	0.15	49
English	0.24	0.23	0.24	0.02	0.03	26
Math	0.47	0.34	0.54	0.20	0.24	14
OR (*overall*)	4.28	4.27	4.30	0.04	0.10	148
Instruction	4.15	4.12	4.18	0.05	0.13	142
Environment	4.64	4.63	4.65	0.02	0.05	142
Planning	4.20	4.18	4.21	0.03	0.05	142
Professionalism	4.52	4.51	4.53	0.02	0.04	146
Questioning	3.99	4.01	3.96	-0.05	0.09	142
Years of experience	13.2	13.5	12.9	-0.57	0.06	148
Total salary	$47,320	$47,413	$47,230	-$183.28	0.02	150
Age	42.3	42.8	41.8	-0.93	0.09	147
White	0.97	0.99	0.95	-0.04	0.21	146
Female	0.88	0.90	0.86	-0.05	0.14	147
Master's degree or higher	0.62	0.72	0.53	-0.19	0.39*	148
Previously mentored	0.80	0.78	0.80	0.01	0.04	113
Previously received PD	0.26	0.36	0.18	-0.19	0.43*	104
Panel C: Program Characteristics						
Year 1	0.33	0.33	0.32	-0.01	0.02	166

Year 2	0.27	0.28	0.25	-0.03	0.06	166
Year 3	0.41	0.39	0.43	0.03	0.07	166
Program 1	0.20	0.20	0.21	0.00	0.01	166
Program 2	0.13	0.14	0.13	-0.013	0.04	166
Program 3	0.07	0.08	0.06	-0.02	0.07	166
Program 4	0.19	0.19	0.18	-0.01	0.02	166
Program 5	0.22	0.22	0.23	0.01	0.04	166
Program 6	0.19	0.18	0.20	0.02	0.05	166
Number of placements	1.20	1.20	1.20	-0.01	0.02	166
Yearlong placement	0.48	0.47	0.48	0.01	0.03	166

Note. Joint chi-squared test for 166 observations with 37 degrees of freedom = 45.03, $p = .17$. When candidates had more than one mentor, continuous mentor characteristics (e.g., salary) are calculated as averages and binary characteristics (e.g., white) as maxima (i.e., if-ever). Coverage/missingness patterns do not differ significantly by treatment condition. $^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix Table 2. Codebook

Code	Description
Coaching Moves	
Feedback – Perspective	Subjective; based on inferences and judgments; gives insight, reinforcement/guidance; incorporates knowledge/wisdom of mentor
Feedback – Data-Driven	Mentor uses evidence from an observed lesson to provide feedback to the candidate
Probing Questions	Mentor asks questions to elicit comments and feedback from candidate
Connecting Candidate and Student Actions	Teaching of candidate is discussed in conjunction with the actions of a student or group of students
Identified Areas of Improvement	Mentor or candidate identifies new or repeated area of improvement
Recommendation	Speaker makes suggestions about future practices of candidate
Mentor Practice	Speaker discusses their own practice when providing feedback to candidate
Coaching Content	
Instruction	Standards and Objectives, Motivating Students, Presenting Instructional Content, Lesson Structure and Pacing, Activities and Materials, Questioning, Feedback, Grouping Students, Teacher Content Knowledge, Thinking, Problem-Solving
Questioning	Topic of conversation is specific to the practice of questioning as defined by the TEAM rubric; intended to capture the emphasis of the Mentors Matter Professional Development
Planning	Instructional Plans, Student Work, Assessment
Environment	Expectations, Managing Student Behavior, Environment, Respectful Culture

Appendix Table 3. Coaching Conversation Code Frequencies

Codes	Frequency*	Mentor**	Candidate#
Coaching Moves			
Probing Questions	0.245	0.447	
Feedback – Perspective	0.172	0.313	
Feedback – Data-Driven	0.077	0.157	
Connecting Candidate and Student Actions	0.083	0.071	0.118
Identified Area of Improvement	0.123	0.080	0.174
Recommendation	0.201	0.170	0.237
Mentor Practice	0.034	0.052	0.011
Coaching Content (TEAM Rubric)			
Instruction	0.836	0.845	0.819
Questioning	0.263	0.245	0.251
Planning	0.063	0.054	0.065
Environment	0.042	0.033	0.048

Note. * Number of times code was applied divided by all units of text.
** Number of times code was applied divided by all units of text with mentor speaking.
# Number of times code was applied divided by all units of text with candidate speaking.

Technical Appendix

Technical Appendix Table 1. Measurement Model Factor Names and Alpha Scores

Factor			Alpha	IIC
PRE-	Feelings of Preparedness	Readiness in Questioning Skills	0.886	0.610
		Readiness in Other Instructional Skills	0.884	0.560
POST-	Coaching Frequency	Common Coaching Practices	0.842	0.727
		Data-Driven Coaching Practices	0.931	0.693
		Collaborative Coaching Practices	0.838	0.721
		Modeling Coaching Practices	0.771	0.627
	Coaching Satisfaction	Support and Feedback	0.967	0.767
		Autonomy and Encouragement	0.921	0.745
	Feelings of Preparedness	Readiness in Questioning Skills	0.867	0.619
		Readiness in Other Instructional Skills	0.878	0.591

Technical Appendix Table 2. Measurement Model Fit Indices

Factor		Chi-Squared			RMSEA					
		Value	df	*p*	LB	Est.	UB	CFI	TLI	SRMR
PRE-	Feelings of Prep.	118.22	42	<0.001	0.035	0.065	0.095	0.967	0.957	0.038
POST-	Coaching Freq.	69.634	45	0.011	0.044	0.061	0.079	0.988	0.982	0.044
	Coaching Satis.	126.12	61	<0.001	0.084	0.098	0.112	0.972	0.963	0.047
	Feelings of Prep.	46.923	25	0.005	0.057	0.082	0.108	0.977	0.967	0.045

Note. All factor scores are standardized for ease of interpretation. RMSEA is root mean squared error of approximation, LB and UB are lower and upper bounds for a 90% confidence interval of RMSEA estimate, CFI is Comparative Fit Index, TLI is Tucker-Lewis Index, and SRMR is standardized root mean square residual. Chi-squared values, CFI, and TLI all obtained using Satorra-Bentler adjustment for non-normal data.

Technical Appendix Table 3. Measurement Model Factor Structure and Loadings

Item	Loading	S.E.
Panel A: Candidate Pre-Survey – Feelings of Preparedness		
Sub-Factor: Preparedness in Questioning Skills		
b. Plan question sequences that help students develop deep conceptual understanding	0.724***	(0.027)
e. Ask questions that require students to discuss and/or write out their developing thoughts	0.777***	(0.022)
i. Develop questions that prompt students to grapple with the elements most necessary for understanding a text or concept	0.814***	(0.019)
j. Challenge students to wrestle with deep questions by providing adequate wait time	0.803***	(0.020)
k. Challenge all students by using strategies for calling on all students equitably	0.789***	(0.021)
Sub-Factor: Preparedness in Other Instruction Skills		
a. Focus on essential information when presenting content	0.713***	(0.025)
c. Provide activities and materials that are relevant to students' lives	0.791***	(0.023)
d. Provide examples, illustrations, analogies, and labels for new concepts and ideas	0.781***	(0.025)
f. Plan activities that build curiosity	0.752***	(0.025)
g. Present content using visuals that establish the purpose of the lesson	0.830***	(0.021)
h. Incorporate multimedia, technology, and resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.)	0.688***	(0.035)
Covariance Structure		
Questioning with Other	0.909***	(0.044)
Residual for item *a* with residual for item *b*	0.315***	(0.044)
Panel B: Candidate Post-Survey – Coaching Frequency		
Sub-Factor: Common Coaching Practices		
a. Observe you teach	0.824***	(0.028)
c. Prompt you to practice specific aspects of teaching during a lesson	0.847***	(0.028)
Sub-Factor: Data-Driven Coaching Practices		
i. Share data or evidence about lessons s/he observed you teach	0.836***	(0.022)
j. Ask you reflective questions	0.834***	(0.023)
k. Analyze student work with you	0.758***	(0.029)
m. Use evaluation data to provide recommendations for improvement	0.867***	(0.017)
n. Provide opportunities outside of regular instruction to practice specific teaching moves	0.760***	(0.032)
l. Share specific next steps for you to work on in order to improve your teaching	0.873***	(0.022)
Sub-Factor: Collaborative Coaching Practices		

d. Co-plan a lesson or activity with you	0.957*** (0.027)
e. Co-teach a lesson or part of a lesson with you	0.753*** (0.035)
Sub-Factor: Modeling Coaching Practices	
g. Model a specific instructional skill or move when students were not present	0.844*** (0.032)
h. Model a specific instructional skill or move for you during a lesson	0.741*** (0.041)
Covariance Structure	
Common with Data-Driven	0.896*** (0.028)
Common with Collaborative	0.708*** (0.050)
Common with Modeling	0.793*** (0.047)
Data-Driven with Collaborative	0.700*** (0.044)
Data-Driven with Modeling	0.834*** (0.040)
Collaborative with Modeling	0.599*** (0.056)
Residual for item m with residual for item n	0.274*** (0.058)
Residual for item m with residual for item l	0.221*** (0.077)
Residual for item e with residual for item h	0.344*** (0.063)
Panel C: Candidate Post-Survey – Coaching Satisfaction	
Sub-Factor: Support and Feedback	
a. My clinical mentor helped me identify next steps to improve my teaching.	0.871*** (0.035)
d. My clinical mentor provided helpful coaching about presenting instructional content that helped me improve my teaching.	0.933*** (0.012)
e. My clinical mentor provided helpful coaching about planning instructional activities and materials that helped me improve my teaching.	0.914*** (0.016)
f. My clinical mentor provided helpful coaching about questioning students about instructional content that helped me improve my teaching.	0.868*** (0.023)
g. My clinical mentor explained how changing certain aspects of my teaching would improve student learning.	0.814*** (0.033)
c. When my clinical mentor observed and evaluated my teaching, I felt her/his evaluations were accurate.	0.809*** (0.033)
h. Overall, my clinical mentor's feedback helped me to improve.	0.904*** (0.018)
i. My clinical mentor observed me teach frequently enough.	0.858*** (0.025)
j. My clinical mentor provided me with feedback frequently enough.	0.905*** (0.016)
Sub-Factor: Autonomy and Encouragement	
k. When I struggled with my teaching, I felt comfortable going to my clinical mentor for help.	0.910*** (0.020)
l. My clinical mentor’s expectations of me were appropriate to my experience.	0.898*** (0.022)
m. My clinical mentor allowed me to make my own instructional decisions.	0.785*** (0.046)
n. I felt comfortable taking instructional risks in front of my clinical mentor.	0.780*** (0.040)
Covariance Structure	
Support and Feedback with Autonomy and Encouragement	0.903*** (0.023)

Residual for item e with residual for item i	0.137 (0.086)
Residual for item f with residual for item g	0.315*** (0.094)
Residual for item m with residual for item n	0.384*** (0.073)
Panel D: Candidate Post-Survey – Feelings of Preparedness	
Sub-Factor: Preparedness in Questioning Skills	
b. Plan question sequences that help students develop deep conceptual understanding	0.794*** (0.036)
i. Develop questions that prompt students to grapple with the elements most necessary for understanding a text or concept	0.848*** (0.023)
j. Challenge students to wrestle with deep questions by providing adequate wait time	0.783*** (0.035)
k. Challenge all students by using strategies for calling on all students equitably	0.709*** (0.045)
Sub-Factor: Preparedness in Other Instructional Skills	
a. Focus on essential information when presenting content	0.672*** (0.044)
c. Provide activities and materials that are relevant to students' lives	0.770*** (0.040)
d. Provide examples, illustrations, analogies, and labels for new concepts and ideas	0.826*** (0.027)
f. Plan activities that build curiosity	0.720*** (0.040)
g. Present content using visuals that establish the purpose of the lesson	0.840*** (0.029)
Covariance Structure	
Questioning with Other	0.827*** (0.033)
Residual for item j with residual for item k	0.202* (0.079)

Note. Robust standard errors. $^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.