Suggested Citation:
Ronfeldt, M., Bardelli, E., Truwit, M., Mullman, H., Schaaf, K., & Baker, J. C. (2020). Improving Preservice Teachers' Feelings of Preparedness to Teach Through Recruitment of Instructionally Effective and Experienced Cooperating Teachers: A Randomized Experiment. Educational Evaluation and Policy Analysis, 42(4), 551–575. https://doi.org/10.3102/0162373720954183

Improving Preservice Teachers' Feelings of Preparedness to Teach Through Recruitment of Instructionally Effective and Experienced Cooperating Teachers: A Randomized Experiment

Matthew Ronfeldt

Emanuele Bardelli

Matthew Truwit

Hannah Mullman

University of Michigan

Kevin Schaaf

Tennessee Department of Education

Julie C. Baker

Tennessee Technological University

Abstract: New studies show that the instructional effectiveness of preservice candidates and their cooperating teachers are positively related. However, we neither know if these relationships are causal nor, assuming they are, if it is possible to significantly increase the instructional effectiveness of the cooperating teacher pool. In this study, we randomly assign districts to receive recommendation lists (generated using administrative data) for the recruitment of more promising cooperating teachers. Districts receiving lists recruited significantly more effective/experienced cooperating teachers, while candidates placed in these districts felt significantly better prepared to teach. As a result, this study offers an innovative, low-cost strategy for recruiting effective/experienced cooperating teachers and presents the first causal estimates that more effective/experienced cooperating teachers improve candidates' preparedness to teach.

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 Hannah Mullman also received pre-doctoral support from the Institute of Education Sciences (IES), U.S. Department of Education (PR/Award R305B150012). We appreciate comments on an earlier draft of this paper from the Tennessee Education Research Alliance, as well as attendees at the 2019 American Education Finance and Policy conference in Kansas City, MO and at the 2019 American Educational Research Association conference in Toronto, ON. We are grateful to Dan Goldhaber and James Cowan for their contributions to the Improving Student Teaching Initiative (ISTI) which established a conceptual and methodological foundation for the present study. We are also grateful to Matt Diemer and Michael Frisby for offering methodological consultation. This project would not have been possible without the partnership, support, and data provided by the Tennessee Department of Education. Any errors should be attributed to the authors.

Matthew Ronfeldt is Associate Professor of Educational Studies at the University of Michigan. His research intends to inform policy and practice by focusing on identifying factors that promote entry into teaching, instructional

quality, and retention, including features of preservice teacher education, school working conditions, and induction supports. Address: 610 E. University Ave., Ann Arbor, MI 48109. Email: ronfeldt@umich.edu.

Emanuele Bardelli is a doctoral candidate in educational studies and a fellow in the causal inference in education policy research predoctoral training program at the University of Michigan School of Education. His research interests include teacher professional development, teacher learning, and instructional practices in mathematics education. Address: 610 E. University Ave., Ann Arbor, MI 48109. Email: bardelli@umich.edu.

Matthew Truwit is a graduate student pursuing his doctorate in quantitative research methods in education and a master's in statistics at the University of Michigan. He seeks to incorporates a comprehensive perspective on the developmental role of schooling with causal inference methodologies to evaluate the broad effectiveness of policies and programs designed to address educational inequity. Address: 610 E. University Ave., Ann Arbor, MI 48109. Email: mtruwit@umich.edu.

Hannah Mullman is a doctoral student in educational studies and a fellow in the causal inference in education policy research predoctoral training program at the University of Michigan School of Education. Her research interests include teacher learning and the relationships teachers form with their students. Address: 610 E. University Ave., Ann Arbor, MI 48109. Email: hmullman@umich.edu.

Kevin Schaaf, Ph.D., is the Director of Research and Evaluation at the Tennessee Department of Education. In this role, he manages a team of researchers and supports the Executive Leadership team at the department in making evidence-based decisions. His research focuses on measuring classroom instruction, teacher preparation, teacher evaluation and development, and the intersections between research and practice. Address: 710 James Robertson Parkway, Nashville, TN 37243. Email: kevin.schaaf@tn.gov.

Julie C. Baker is an Associate Professor in the College of Education at Tennessee Tech University. She currently serves as the Associate Dean in the College, working closely with teacher education, accreditation, undergraduate curriculum, and student success. Her research interests include rural schools and communities, adolescent reading, postsecondary persistence, and preservice teacher education. Address: Campus Box 5016, Cookeville, TN 38505. Email: jcbaker@tntech.edu.

Introduction

In order to receive initial certification, a teacher candidate completes clinical training, often referred to as student teaching or residency, in the classroom of a cooperating teacher (CT) – a P-12 teacher who mentors them as they take on classroom teaching responsibilities. There is increasing evidence that clinical training – and the CT specifically – has important influences on preservice student teacher (PST) development (Bastian et al., 2018; Goldhaber et al., 2018a; Kang, 2020; Ronfeldt, Brockman, et al., 2018). New research suggests that PSTs who were mentored by instructionally effective teachers, as measured by observational ratings of performance or value-added to student achievement models (VAMs), are more instructionally effective themselves once employed (Goldhaber et al., 2018a; Ronfeldt, Brockman, et al., 2018; Ronfeldt, Matsko, et al., 2020). Therefore, as teacher education programs strive to provide the best possible preparation for their candidates, selecting higher rated CTs is a promising lever.

However, program leaders report that it is often difficult to recruit instructionally effective teachers to serve as CTs for a variety of reasons. First, due to privacy and limited availability of evaluation data, school, district and program leaders may not know who the most instructionally effective teachers in local districts are. Second, school and district leaders may be resistant to handing responsibility of instruction over to novice teachers in the classrooms of their strongest teachers. Finally, these different stakeholders may have competing criteria for selecting CTs, some of which may not relate to their instructional effectiveness (Krieg et al., 2020; Ronfeldt, et al., 2020).

This study describes an initiative that aims to increase the overall instructional effectiveness of teachers serving as CTs. In particular, the initiative tests whether providing data-driven recommendations for the targeted recruitment of CTs to district and teacher education

program leaders can both raise the average level of effectiveness of CTs and improve the quality of preparation for PSTs. First, we created an algorithm to identify the most instructionally effective and experienced teachers in the districts, subjects, and grades in which CTs were needed. We then worked with Tennessee Technological University (TTU) – one of the largest providers of teachers in the state – and the many partnering districts in which it places PSTs to randomly assign districts either to receive and use recommendation lists based on this algorithm or to place PSTs as they normally would. We find that districts that were randomly assigned to use these recommendation lists were able to recruit substantially more effective and experienced teachers than other districts (by 0.4-0.7 standard deviation units across measures). Moreover, we observe that PSTs who learned to teach with this group of CTs felt significantly better prepared to teach at the end of their clinical training (by 0.5-0.7 standard deviation units across measures). These findings provide evidence in support of policies, like those in the state of Tennessee where this study takes place, that set minimum requirements for instructional performance and years of experience in order for teachers to serve as CTs. Furthermore, results indicate that a viable way to implement and even enhance such a policy involves leveraging existing information on the instructional effectiveness and experience of teachers to recommend which teachers to target. The findings demonstrate that, even within a state context with policy already establishing minimum evaluation requirements for service, the provision of improved information can meaningfully increase the quality of the pool of CTs.

Background / Literature Review

Understanding CTs' instructional effectiveness and its likely effects on PSTs

Three new studies have come to the same conclusion: PSTs are more instructionally effective early in their careers when they learn to teach with more instructionally effective CTs

during their clinical training. In Tennessee, Ronfeldt, Brockman, and Campbell (2018) linked evaluation data of recent program completers to the evaluation data of their CTs and found that newly hired program completers had better observation ratings based upon the state rubric when their CTs also had better observation ratings; likewise, graduates had better student achievement gains (using TVAAS scores) when their CTs did too. In subsequent studies, Ronfeldt, Matsko, Greene Nolan, and Reininger (2020) found similar patterns between PST and CT observation ratings in Chicago, while Goldhaber, Krieg, and Theobald (2018a) identified comparable associations between PST and CT achievement gains in Washington state. These studies suggest the value of policies like the one in place in Tennessee, which establishes minimum teaching evaluation requirements for teachers to serve as CTs. Finding similar relationships across different studies, labor markets, and sets of measures for instructional effectiveness also suggests that these associations are less likely to reflect spurious correlations than the actual effects of CTs on PSTs. However, all three studies are correlational in nature and thus require subsequent research relying on experimental methods to assure that their results are truly causal, a contribution of the present study.

Additionally, these prior studies provide little guidance as to the possible mechanisms by which CT instructional effectiveness may impact PST instructional effectiveness. Just as important as knowledge of the presence of a relationship between CT and PST instructional effectiveness is an understanding of *how* the former influences the latter. For this guidance, we turn to existing literature reviews on the research in teacher education (and specifically clinical education), which consists of primarily qualitative inquiries, typically self-studies, of individual programs. Based upon Glenn (2006) and prior reviews of the existing research, Grossman, Ronfeldt, and Cohen (2012) suggest that CTs serve at least two major functions: as a model of

teaching and as a mentor or instructional coach who deliberately structures opportunities for new teachers to learn, practice, and receive feedback on their teaching efforts.[1]

Cooperating teachers as models or coaches. Regarding modeling, a number of studies suggest that PSTs learn how to teach, at least in part, from observing their CTs model practice and then emulating that practice. In fact, Koerner, Rust, and Baumgartner (2002) found that the PSTs they surveyed were more likely to classify their CTs as "role models" than "mentors." One might expect, then, that highly effective CTs positively influence PST development by demonstrating best practices that their PSTs are then able to incorporate into their own teaching. Conversely, CTs who model poor instruction may inadvertently pass less effective practices along to their PSTs (e.g., Hoy & Woolfolk, 1990; Rozelle & Wilson, 2012; Zeichner & Gore, 1990); in such a scenario, recruiting instructionally effective CTs could serve as an antidote to some of this detrimental socialization.

CTs may also serve in the role of coaches. Schwille (2008) studied the strategies used by coaches of novice teachers who were known as effective embodiments of "educative mentoring", a practice grounded in learning theories that position the learner (here, the PST) as an active participant in the learning process. Schwille (2008) documented many such strategies – coaching while the PST is in the act of teaching, brief coaching interactions between classes or activities, more formal and structured post-observation debriefs, co-planning and co-teaching lessons, and videotape analysis – in contrast with an "osmosis" approach consistent with modeling, "where the mentor hopes the novice will 'see' and pick up on something on her or his own" (p. 148). In addition to employing different coaching pedagogies, CTs can also provide PSTs with emotional support when needed (Glenn, 2006) and a balance of autonomy and encouragement (Yendol-Hoppey, 2007). While it is likely that more instructionally effective teachers are also more adept

at these coaching practices, it is also possible that they require skills and capacities distinct from those needed for the effective teaching of P-12 students.

The relationship between modeling, coaching, and pre-service teacher outcomes. The empirical basis linking CTs' actual practices – whether as coaches or models – to outcomes for PST learning is especially thin. One exception is McQueen (2018), who designed a training program supporting randomly assigned CTs to provide their PSTs with more choice/autonomy about which area of teaching on which to focus and then to maintain a sustained focus in their feedback on that area over time. Per the typology described above, this training promoted a coaching model for CTs, rather than a modeling one. McQueen found that PSTs who worked with trained CTs received stronger evaluations on their teaching, though differences were significant in only some specifications. These results are also consistent with a large body of research finding consistently positive effects of professional development programs that target the coaching practices of mentors of inservice, rather than preservice, teachers (see Kraft et al., 2018, for a review of this literature).

We know of three other studies that attempt to link the coaching practices of CTs with PST outcomes. Matsko et al. (2020) looked at all CTs who served in the Chicago area and found that PSTs reported feeling better prepared to teach at the end of their programs when also reporting that their CTs provided more frequent and/or a higher quality of feedback, instructional support, autonomy and encouragement, collaborative coaching, and job assistance. In a subsequent, related study, Ronfeldt, Matsko, et al. (2020) also found that PSTs had better first-year observation ratings (based upon district evaluations) when their CTs reported more coaching focused on specific instructional practices, including those evaluated on the same rubric. However, both studies still suggest that the modeling function of CTs also benefits PSTs.

The better prepared PSTs in Matsko et al. (2020) reported more effective instructional modeling from their CTs, who also received better observation ratings (based on the district evaluation rubric), while the more instructionally effective first-year teachers in Ronfeldt, Matsko, et al. (2020) served under CTs with higher observation ratings as well.

The third study provides the most relevant evidence about whether more instructionally effective CTs provide higher quality coaching. Ronfeldt, Goldhaber, et al. (2018) developed and evaluated an initiative similar in many ways to the study described in this paper, combining prior administrative data on CTs' instructional performance and experience with the average student achievement gains and teacher retention rates of the schools in which they worked to create an index for predicting more and less promising placements. Using the median as the cutoff, the authors randomly assigned PSTs to be placed in either low- or high-index placements. Compared to their peers in low-index placements, high-index PSTs reported that their CTs not only modeled more effective instructional practices but also engaged in more frequent and higher quality coaching activities like the provision of feedback and opportunities to practice different aspects of teaching in their placements – in short, both modeling and coaching again.

Though these results offer some plausibly causal estimates for the relationship between CT and PST instructional effectiveness, as well as some suggestion of the mechanisms by which that relationship has impact, other non-causal explanations are still possible. By including school-based measures like average teacher retention, which is known to signal school working conditions (Ronfeldt, 2012), CTs may simply have had more opportunity to mentor PSTs as a function of the characteristics of their schools rather than of their own attributes. Our present study rules out this alternative explanation by focusing only on measures of teachers' instructional effectiveness and experience absent any school-level variables; it thus offers the

best causal evidence to date for the impact of being assigned to an instructionally effective and experienced CT on PST preparedness to teach.

Another limitation of Ronfeldt, Goldhaber, et al. (2018) is that the comparison group in this previous study was, in some sense, manufactured as a part of the research, with control PSTs intentionally assigned less promising placements from the lower half of the index. In contrast, the design of the present study – randomizing at the district level in order to obtain a business-as-usual comparison – allows us to test whether an intervention that is relatively low-cost and easy to reproduce can improve CT recruitment procedures, on average, over typical approaches.

The main purpose of the present study, then, is to use existing administrative data to identify the most instructionally effective and experienced teachers to serve as CTs and then to randomly assign districts to receive recommendation lists (based on this information) to target their recruitment. We then investigate the effects of having an instructionally effective CT on PSTs. Finally, we explore evidence related to two possible mechanisms by which instructionally effective CTs might influence PSTs' preparedness: (1) modeling better instruction or (2) providing better coaching and feedback. The following questions guide this study:

RQ 1. Do CTs in districts randomized to receive recommendation lists have higher average effectiveness scores and experience compared to those in districts following business-as-usual recruitment strategies?

RQ 2. Do PSTs report feeling more instructionally prepared when their CTs were recruited using recommendation lists?

RQ 3. Do PSTs report more frequent and/or higher quality coaching practices when their CTs were recruited using recommendation lists?

What We Know About Recruitment Procedures

The review above illustrates that there is already substantial evidence that recruiting instructionally effective and experienced teachers to serve as CTs is likely a good idea. To what degree is this already a priority among program and district/school leaders? In this section we review the existing literature about how CTs are recruited for student teaching placements, the kinds of obstacles that program and district/school leaders face in recruiting teachers (especially instructionally effective ones) to serve as CTs, and whether or not existing recruitment procedures are already targeting and getting the most instructionally effective teachers to serve.

Existing recruitment procedures. A handful of empirical studies help shed light on the factors that influence the selection of CTs. In particular, demographic match between PST and CT, proximity to the teacher education program (TEP), and CT and placement school characteristics seem to influence which teachers get selected to serve (Krieg et al., 2016; Maier & Youngs, 2009). We know of two studies that explored the recruitment procedures in specific labor markets. Reflecting the literature reviewed above, both of these studies found that TEP leaders and other stakeholders report considering a potential CT's ability not only to model effective instruction with students but also to support and coach a PST. In the first of these studies, St. John, Goldhaber, and Krieg (2018) identified a common CT recruitment process used across eight TEPs in Washington state. Broadly, TEPs began by assessing their needs and contacting district and schools. Schools and districts evaluated their capacity to host, and eventually, PSTs, CTs, and principals met to determine whether each placement was a good match. The authors pointed out that the day-to-day demands and concerns of different stakeholders may cause recruitment procedures to deviate from these steps, but that, for the most part, TEPs adhered to a largely uniform process.

Conversely, in Tennessee, where the present study takes place, Mullman and Ronfeldt (2019) found that recruitment procedures varied both across and within TEPs. Districts and schools each assumed different roles and responsibilities for the selection of CTs along a spectrum ranging from maintaining full control of the process to allowing PSTs themselves to find their own placements. Moreover, if a single TEP placed PSTs in multiple districts, it often used a variety of systems for selecting CTs. Tennessee TEPs also face additional considerations for selecting CTs due to state policies put in place for clinical practice, including requirements for diversity of experience and a minimum of two clinical placements.

In their National Council on Teacher Quality (NCTQ) report, Rickenbrode, Drake, Pomerance, and Walsh (2018) looked across the TEPs of graduate students and concluded that CTs' instructional effectiveness was not a consistent priority in recruitment procedures. Out of the 506 TEPs they studied, they found that even in the eight states that set effectiveness criteria for CTs, only about half of programs took action to ensure these were honored and met. And in the context of discussions of the current study, stakeholders raised a variety of potentially competing priorities that might play a role in recruitment, including rewarding seniority, providing "help" to a struggling teacher, and practicing turn-taking to give every teacher a chance to serve as a CT.

Challenges to recruiting (instructionally effective) CTs. Both St. John and colleagues (2018) and Mullman and Ronfeldt (2019) identified knowledge gaps as obstacles to recruiting instructionally effective CTs. Typically, due to privacy laws, data about value-added to student achievement and observation ratings are not available to TEPs or PSTs. Even in contexts where evaluation data is accessible, discrete categorizations and/or a highly compressed and top-heavy distribution of teacher quality can make it challenging to differentiate among higher performing

teachers despite meaningful variation in their effectiveness. Furthermore, even when district leaders and school administrators might know who the most effective teachers are, they may not share that information with TEPs. Mullman and Ronfeldt (2019) talked to TEP leaders who said they simply had to trust that their district partners were complying with state regulations for instructional effectiveness. Additionally, both studies described above found evidence that stakeholders may prioritize other traits when selecting CTs. These included differences in opinion about the role of the CT (i.e., as coach or model), social networks (TEPs often recruit alumni from their programs to serve), and ease of onboarding (given that once a TEP has a relationship with a teacher, they may try to use that CT again). There is also a prevalent belief that working with adult learners differs from working with young learners, so instructional effectiveness measures might not tell TEPs a great deal about a teacher's capacity to mentor a PST (St. John et al., 2018; Mullman & Ronfeldt, 2019).

It is also plausible that the most effective teachers may be hesitant to serve in the current climate of accountability. Teachers who serve as CTs give a large portion of instructional time to their less experienced PSTs, which they fear may negatively impact their value-added scores (Goldhaber et al., 2018a; Ronfeldt, et al., 2020; SAS Institute, 2014). In Tennessee, Ronfeldt and colleagues (2018) explored this possibility by measuring the impact of serving as a CT on both value-added measures and observation ratings. While they allay these concerns, finding no effects on value-added and small, positive effects on observation ratings, hesitancy to serve on the part of teachers may remain. St. John and colleagues (2018) also found concerns that CTs who served multiple times might feel burnout. Mentoring a novice requires a great deal of time and effort, and the work is rarely compensated more than a few hundred dollars, if at all. Some

stakeholders interviewed by the authors reported feeling reluctant to ask the same high-quality CTs to serve repeatedly as they worried about putting undue burden on these teachers.

Do existing recruitment procedures work? The wide variation in recruitment procedures – as well as the many obstacles to recruitment– cast some doubt that existing practices always result in selection of the most instructionally effective CTs. Yet, there is some evidence that program and district/school leaders are already recruiting individuals to serve as CTs who are relatively more effective and experienced than other teachers. Examining 21 programs in Tennessee, for example, Ronfeldt, Brockman, and Campbell (2018) found that CTs had significantly better observation ratings and VAMs than other teachers in the state, though they had similar levels of teaching experience. Across preparation programs in Chicago, Gordon et al. (2018) found that, compared with their peers who did not serve as mentors, CTs had better REACH observation ratings and were more likely to have a master's degree, be tenured, and be National Board Certified; however, they had statistically similar VAM scores. In Washington state, Goldhaber et al. (2018b) discovered that, all else being equal, teachers with more experience were more likely to host a PST, but teachers with greater VAMs were not.

Given that recruiters seemed to already be tapping more instructionally effective and experienced teachers to serve as CTs, we were concerned that the pool of effective teachers in needed grades/subjects/districts might already be exhausted. If so, then supplying district/school/program leaders with recommendations about effective and experienced teachers to recruit might have little or no effect. We wondered whether there would even be enough alternative, more effective teachers willing to serve to make a significant difference, and if so, we were concerned that some of the other obstacles described above might obstruct efforts to use recommendation lists to nudge recruitment. As a result, our first research question centers on

whether or not providing recommendation lists alone increases the effectiveness and experience of recruited CTs, while our second and third return to the issue of whether and how instructionally effective CTs impact PST preparation.

It is also instructive to emphasize here the strong policy relevance of this first research question. This study was designed in close collaboration with our state department partners in an effort to provide a test of the lowest cost policy lever that we identified as a means of potentially raising the overall instructional effectiveness of the pool of CTs. Initial study design conversations considered the possibility of testing the use of cash incentives as a means of attempting to recruit more instructionally effective teachers to serve as CTs, but that idea was shelved in favor of the present study out of concern for the need to test a strategy that could be sustainable and scalable in the absence of grant funds. Moreover, jumping right to incentives would have presumed that providing better recruitment information (absent accompanying incentives) would not suffice, so we decided to test whether providing better information alone could move the needle before adopting incentives.

Method

Research Design, Context, and Sample

For this initiative, we partnered with TTU, a large provider that uses a residency model, where PSTs complete a year-long clinical placement[2] in their CTs'[3] classroom(s). In 2017-18, the program placed 189 PSTs in 22 neighboring districts. PSTs needed to complete their residency in subjects/grade levels appropriate for their specific program endorsement areas; for example, those pursuing elementary endorsements were placed in grades K-5. Additionally, PSTs were able to request a specific county/district in which they wanted to be placed, especially to accommodate geographic and travel constraints.[4] We report PST pre-recruitment characteristics

in Appendix Table 1. The majority of PSTs identified as being White and female, and on average, they had a 3.46 GPA, an admission ACT score of 22.80, and a Praxis score of 168.94. About 63% of the teachers requested an elementary education clinical placement. Seventeen PSTs left the TEP during the duration of our experiment, translating to an overall attrition rate of about 9%. We do not find differences in PST attrition by treatment condition.

We used clinical placement request information to identify, for each PST, all teachers that matched the county/district-by-grade band-by-subject "block" of choice. We then used prior information on instructional performance and years of experience (from administrative data) to identify the most instructionally effective and experienced potential CTs in these blocks (see below for details) and – based upon this information – generated recommendation lists to guide CT recruitment. Table 1 reports summary statistics for these potential teachers. Overall, teachers who were selected to serve as CTs, regardless of the treatment condition, appear to have higher evaluation scores and years of experience than other potential CTs who work in a similar assignment in the same county/district, which aligns with the findings of previous work.

Our state partners, with our technical support, then randomly assigned neighboring districts to receive these recommendation lists and requested district leaders who received the lists to use them in their recruitment, starting where possible with the teacher at the top of the list (highest ranked). District leaders were also advised to use their best judgement and to skip any listed teachers that they felt were inappropriate or unwise to recruit and to instead move to the next listed teachers. We presumed that providing district leaders with this flexibility would both honor and leverage their personal expertise and their knowledge of the strengths of their teaching staffs. Among districts randomly assigned to treatment, district leaders took primary responsibility for outreach and recruitment in ten districts, while TTU leaders took primary

responsibility in two districts; in the latter case, TTU leaders reached out directly to school leaders and/or specific teachers. In these cases, the state shared recommendation lists with TTU leaders, who then used them for CT recruitment.

We asked that recruiters (district or program leaders) who received the recommendation lists keep notes on which teachers were invited, accepted invitations to serve, and declined; for the latter, we requested that recruiters record and share notes on why teachers declined. We received data from twelve treatment districts and 160 teachers who were contacted during the recruitment drive; 92 teachers (55.4%) accepted to serve as a CT and 74 teachers (44.6%) declined to serve when offered. Among those that declined, recruiters entered notes on why for 60 of these teachers. Many teachers (n=17, or 28.3%) declined to serve for personal reasons, including, for example, not being interested in serving, not having enough time to properly supervise a new teacher, or wanting to serve only for a semester instead of a yearlong clinical placement. District administrators declined the recruitment request for 11 teachers (or 18.3%). Our lists had the wrong information for 26 teachers; 16 of these potential CTs (or 26.7%) had either left their teaching assignment or had their teaching assignment misidentified (10 teachers or 16.7%). Other reasons were given for 6 teachers (or 10%). In all cases where teachers declined to serve as CTs, recruiters simply continued to the next name on the list.

Balance Check

Given that randomization occurred at the district level, we report the results of balance checks for observed district characteristics in Table 2. We checked for balance on K-12 student characteristics and potential CT evaluation scores at the district level. There does not appear to be any evidence that the treated and control districts are significantly different across all tested

dimensions. This result suggests that randomization was successful in balancing treatment and control districts on observed teacher and student covariates.

As PST surveys provide the outcomes of interest, we also tested for balance on observed PST demographic characteristics, prior achievement, and Praxis scores at time of randomization. We report these results in Appendix Table 1 to evaluate whether the differential sorting of PSTs into districts might bias any estimates we derive. We find no significant differences in these covariates between PSTs in treatment and control districts, suggesting that randomization was also successful at minimizing observed PST differences between conditions.

Recruitment Procedures

In this section, we elaborate on the specific algorithm used to generate the recommendation lists. We calculated a composite measure for our "recruitment index" as the weighted average of observation ratings[5] (OR), value-added measures[6] (VAMs), and years of experience. We first standardized each measure within recruitment field[7] at the state level. This procedure used the following formula:

$$Y_{STD_i} = \frac{Y_i - \overline{Y_b}}{\sigma_{Y_b}}$$

where $\overline{Y_b}$ and σ_{Y_b} are the state-wide mean and standard deviation for variable Y_i within recruitment block b. For OR and VAM, we averaged the scores for the three preceding school years, weighing the year immediately preceding recruitment as 50% of that measure and the other two 25% each.[8] This can be represented as:

$$EVAL_i = 0.25 \cdot EVAL_{it-3} + 0.25 \cdot EVAL_{it-2} + 0.50 \cdot EVAL_{it-1}$$

We then calculated a final recruitment index as

$$RI_i = 0.40 \cdot OR_i + 0.40 \cdot TVAAS_i + 0.20 \cdot EXP_i$$

where OR_i, $TVAAS_i$, and EXP_i are the standardized weighed averages described above. EXP_i is the number of years of experience reported for school year 2016-17.

Missing evaluation data. We had some missing evaluation data for the years that we used to calculate the recruitment index. We decided not to impute or otherwise calculate possible values for these data. Instead, we just removed the variable from the calculations and adjusted the weights to reflect the data that were present. For example, if observation scores were not available for teacher i for time $t-3$, her evaluation scores were weighed as 0.50 for $t-2$ and 0.50 for $t-1$. Other combinations of missing data followed the same procedure.

We also made the decision to exclude (i.e., treat as missing) individuals for whom only experience, without other quality measures, was available. Our partners at the Tennessee Department of Education have argued that calculating quality based only on years of experience does not add anything new for school administrators and district leaders, as experience is an easy variable to observe in teachers.

CT eligibility. We decided that teachers were eligible to be recommended as CTs when they fell in the upper three quintiles of the recruitment index distribution. Thus, the recommendation lists are organized by recruitment index score, where the potential CTs with the highest index score were at the top of the list and thereby the ones we asked district/TEP leaders to recruit first. If a district/TEP leader exhausted all teachers on the lists and still could not recruit a CT for a PST, at that point, we expected them to recruit in whatever way they typically would otherwise. Our rationale was that their business-as-usual approaches were preferable to suggesting they recruit CTs towards the bottom of the index distribution. Additionally, we wanted to avoid the possibility of recommending CTs who may not have met the minimum level of effectiveness (LOE) score to serve as a CT.

This approach also mitigated a potential sensitivity that might have arisen with the practice of sending districts a ranked list of recommended teachers. Because our list only included teachers who were ranked approximately at or above "average" on our recruitment index, we hoped to assuage any concerns district partners might have about receiving "ranked" lists of teachers. That is, although potential CTs were still ordered from most instructionally effective down, all teachers on the list were recommended, and consequently there should have been no stigma associated with appearing near the bottom of the list.

Outcomes of Interest

Along with the selected CTs' rankings based on value-added, observation, and experience, the outcomes of interest for this paper include survey-based reports of feelings of preparedness, frequency of coaching, and satisfaction with coaching. We surveyed PSTs at the beginning (pre-survey) and at the end (post-survey) of their clinical placement. We also surveyed CTs once during the second half of the clinical placement. PSTs were surveyed about all outcomes, while CTs were asked to report on the frequency of their coaching. The survey items were adapted from instruments used previously (Ronfeldt, Goldhaber, et al., 2018; Matsko et al., 2020) to assess the preparation of student teachers. We adapted these prior instruments to this initiative, developing new items to collect data on the goals of the experiment and to better align with the teacher evaluation system in Tennessee. We report non-response rates for each survey in Appendix Table 2. The control group response rate was 41.6% for the PST post-placement survey and 71.0% for the CT survey. We do not find evidence of differential non-response by treatment condition.

In the following sections, we provide a qualitative description of each latent construct we include in our analyses. Technical Appendix 1 reports in detail the psychometric procedures we

followed to calculate factor scores for each measure, including reliability estimates, fit indices, and factor loadings for each model.

Feelings of preparedness (PST survey). We measured feelings of preparedness in both pre- and post-surveys. We divided this construct into two correlated sub-constructs: preparedness in questioning skills and in other instructional skills. The first sub-factor includes five items focused on preparedness in developing, planning, and implementing questions to engage students in understanding a concept; we included a focus on this construct because the state had identified this as a priority, especially since "questioning" is consistently amongst the lowest rated indicators, on average, on the TEAM rubric across the state. The second sub-factor includes six items about other aspects of planning and delivering instruction, such as developing materials, providing examples or analogies for new concepts, and using visuals during a lesson.

Coaching frequency (PST survey). We measured frequency of coaching practices using four sub-constructs that focus on common coaching practices, data-driven coaching practices, collaborative coaching practices, and modeling coaching practices. Common coaching practices include two items asking about the frequency of observations and of prompts to practice a specific aspect of teaching practice. We have seen these coaching practices to be the most commonly used during student teaching and therefore practices with which all CTs are likely familiar. Data-driven coaching practices include six items that focus on using data from observations or student work to guide coaching. Collaborative coaching includes two items focused on co-planning and co-teaching activities, while modeling coaching practices include two items assessing modeling of specific instructional strategies by the CT.

Coaching satisfaction (PST survey). We measured coaching satisfaction using two sub-constructs that include support/feedback and autonomy/encouragement. The support and

feedback sub-factor includes nine items that measure satisfaction with specific coaching practices (i.e., identifying next steps to improve teaching; coaching about instructional content, planning instructional activities, and questioning students; explaining how certain changes to practice would impact student learning) and feedback (i.e., feeling that their CTs' evaluations and feedback were accurate, helpful, and sufficiently frequent). The autonomy and encouragement sub-factor includes four items that measure the extent to which PSTs felt comfortable and independent under their CTs (i.e., feeling comfortable asking their CTs for help and taking risks in front of them, feeling that their CTs' expectations were appropriate, and feeling that they had the ability to make independent instructional decisions).

Coaching frequency (CT survey). The CT survey included two main factors for coaching practices: a general factor with three sub-factors and a specific factor on instructional practices. We divided the general factor on frequency of coaching practices into three correlated sub-factors: debriefing, developing practice, and collaborative coaching practices. The debriefing sub-factor includes five items that focus on helping the PST debrief a lesson through questioning, analysis of student work, or data analysis. The developing practice sub-factor includes four items that focus on modeling specific instructional skills or providing opportunities to practice outside of regular instruction. The collaborative coaching practice includes two items measuring the frequency of co-teaching and co-planning activities.

The specific factor includes questions about frequency of coaching around key instructional practices. This factor includes eleven items that are aligned with the instruction domain in the TEAM observation rubric used in Tennessee. We used text from the domain descriptors from the TEAM rubric as question stems for this factor.

For each of the above four constructs (feelings of preparedness, coaching satisfaction, PST coaching frequency, CT coaching frequency), we also average the sub-construct measures to construct "general" measures. For example, we construct a "general" feelings of preparedness measure by averaging scores for the two subconstructs: (1) preparedness in "questioning skills" and (2) preparedness in "other instructional skills."

Analysis

Our experimental design allows us to conduct a relatively simple analysis. In detail, we use linear regression with fixed effects:

$$Y_{ijd} = \beta_0 + \beta_1 \cdot Treat_d + \phi_j + \epsilon_{ijd}$$

where Y_{ijd} is the outcome of interest for CT or PST i in request field j, $Treat_d$ is an indicator variable taking the value of 1 if district d was randomized to receive a recommendation list, ϕ_j is a recruitment field fixed effect, and ϵ_{ijd} are standard errors clustered at the district level. β_1 captures the treatment effect of receiving the recommendation list on the outcome of interest.

We re-specify our preferred model in three alternative specifications. First, we calculate standard errors using a bootstrap procedure. This allows us to calculate standard errors using a non-parametric, data-driven procedure that might be more robust against violation of the assumptions of our preferred models. Second, we use recruitment field random effects, rather than a fixed effects approach.[9] Though we prefer the fixed effects specification because it adjusts for unobserved differences between recruitment field types, the random effects approach appears to be more efficient.[10] Third, when using feelings of preparedness as an outcome, we include pre-placement feeling of preparedness scores as a covariate in our models in order to control for possible imbalance in PSTs' initial feelings of preparedness (see Appendix Table 3). Overall, we

find that the results of our preferred models are similar to those from alternative specifications; therefore, we mostly discuss results from our fixed effects models from this point forward.

Mediation Analysis

We conduct three mediation analyses to decompose our observed treatment effect into multiple possible pathways that help us explore possible treatment mechanisms, as shown in Figure 1. In each of these structural equation models, treatment influences PSTs' feelings of preparedness through three paths. One indirect path – the "coaching" path – estimates the effect of a district receiving recommendation lists through various survey measures of CTs' coaching capacity: in Panel A, CTs' reported frequency of coaching, in Panel B, PSTs' reported frequency of coaching received, and in Panel C, PSTs' satisfaction with coaching received. Meanwhile, the "modeling" path estimates the indirect effect of treatment through a measure of CTs' instructional effectiveness, using the recruitment index calculated to generate recommendation lists (see Recruitment Procedure in Method for more details). After isolating the potentially positive influence of more instructionally effective CTs' coaching into a separate pathway, this channel essentially proxies only the effect of CT modeling on PST preparedness. The final path directly connects the treatment indicator to the outcome. This pathway contains not only any other channels through which treatment might have influenced PST preparedness but also any effect of CT coaching not contained by our survey measures and modeling not included in our recruitment index.

Each of our operationalizations of CT modeling and coaching likely do not fully capture the constructs that they purport to measure; as a result, much of the treatment effect in each mediation analysis remains in the direct path between treatment and preparedness. As we are inherently unable to determine whether this indicates shortcomings of our measures for modeling

and coaching or an alternative channel through which treatment affects PST preparedness, we choose to avoid interpretation of this pathway. However, under certain strict assumptions, we can interpret each indirect channel relative to the other. Given that any path coefficients for measures with imperfect reliability will be biased toward zero (Bollen, 1989, pp. 154–159), if we assume equivalent reliability for the measures of both constructs, we can assess the extent to which CT modeling or coaching explains more of the observed treatment effect. That said, we acknowledge that our measures for CT coaching and modeling are limited in that they likely fail to capture important aspects of both constructs. Moreover, it is possible that our measures overlap to some degree, where our measure for coaching may capture dimensions of modeling and vice versa. We try to address this concern by correlating the residual terms for the two measures. These correlation terms had small non-significant estimates, indicating that our measures of coaching and modeling do in fact capture non-overlapping aspects of the clinical experience. Regardless, these mediation analyses should be considered as largely exploratory and descriptive, aiming to provide a conceptual and methodological foundation for future research aiming to disentangle modeling and coaching effects.

Results

RQ 1. Recruitment Index Contrast

Table 3 summarizes the differences between CTs in districts randomized to receive recommendation lists (treatment) and districts that use business-as-usual recruitment procedures (control). Overall, we find that CTs in treatment districts have, on average, higher evaluation scores than CTs in control districts. These differences are significant on observation ratings (0.332 s.d. units), VAM scores (0.654 s.d. units), and years of experience (0.558 s.d. units).

We add indicators for recruitment field (i.e., district by grade by subject) requests to increase the statistical power of these analyses and to account for possible differences between recruitment fields such as the possibility that secondary English language arts teachers are rated higher (or lower) on average than, say, elementary teachers. The estimates for these models are reported in the fourth row of Table 3. We find that the point estimates increase slightly, indicating that there are differences on evaluation scores between recruitment fields.

We use the average recruitment index to calculate the overall contrast between treatment and control CTs. This index allows us to compare CTs across recruitment fields as this variable is standardized within each. We find that the recruitment index for CTs in treatment districts is 0.425 standard deviations higher than it is for CTs in control districts. This result is statistically significant at the 0.01 level. When we adjust these estimates for recruitment field differences,[11] we find that the quality contrast increases to 0.476 standard deviation units.[12]

RQ 2. Feelings of Preparedness

Table 4 reports the effect of being placed in a district that received the recommendation lists on PSTs' feelings of preparedness. We find that PSTs in treatment districts reported feeling significantly better prepared to teach by 0.593 standard deviation units (s.e. = 0.226, $p < 0.05$). We also see that these results are robust to how we calculate standard errors, to the inclusion of pre-recruitment controls, and to the re-specification of the model using random effects.

When we focus on the feelings of preparedness in specific subskills, we find that PSTs in treatment districts reported feeling better prepared in both questioning skills ($d = 0.637$, s.e. = 0.230, $p < 0.05$) and other instructional skills ($d = 0.548$, s.e. = 0.225, $p < 0.05$), suggesting that the treatment effect was equally distributed across all teaching sub-skills that we measured.

RQ 3. Reported Coaching

Finding that PSTs in treatment districts felt better prepared made us wonder about the underlying mechanisms driving these differences. One potential explanation is that, by depending upon the recommendation lists, these districts recruited more effective and experienced teachers to serve as CTs; in turn, perhaps more effective and experienced teachers, on average, model better instruction, thus helping PSTs to feel better prepared by regularly observing best practices. Another possibility is that more effective and experienced CTs, on average, provide more or better instructional coaching to their PSTs. To test this second possibility, we examined survey items related to the frequency of and satisfaction with coaching that PSTs reported receiving and that CTs reported offering.

Results, which are summarized in Table 5, suggest that PSTs in treatment districts felt they received somewhat more frequent coaching activities, as coefficients trend positive across outcomes and model specification; however, results are mostly non-significant. Effects are largest in magnitude (about 0.20 standard deviation units) in relation to data-driven coaching practices. On the other hand, PSTs in treatment districts tended to report less support and satisfaction with the coaching they received and less autonomy and encouragement, though, again, not at significant levels. In terms of the coaching activities that CTs themselves reported, differences between conditions are also mostly non-significant. That said, there were some notable trends: treatment CTs reported engaging in debriefing practices and coaching focused on the "instruction" domain more often and in developing practices less often than control CTs.

Figure 1 and Appendix Table 4 report the results of descriptive mediation analyses that explore the possible mechanisms through which treatment assignment might impact PSTs' feelings of preparedness. The results of these analyses should not be interpreted as causal

relationships between the mediator variables and PSTs' feelings of preparedness as our experimental design only allows us to estimate the causal link between treatment assignment and downstream outcomes. We note three main findings from the results of these mediation analyses. First, focusing on the modeling pathway, our recruitment index appears to explain between twenty-five and twenty-eight percent of the overall treatment effect of receiving a recommendation list on PSTs' feelings of preparedness. This estimate is consistent across the three models regardless of the measure of coaching that we include in our path diagram. Second, coaching frequency appears to explain at most twelve percent of the total treatment effect. The higher end of this range stems from PSTs' reports of coaching frequency, which appear to explain more of the treatment effect than the self-reports of CTs, perhaps suggesting that PSTs are more reliable in reporting the frequency of coaching received than CTs themselves. This intuition is in line with prior work that has found that CTs tend to overreport the frequency of their own coaching practices (Matsko et al., 2020). Third, we find that PSTs' satisfaction with coaching could actually work as a suppressor of our treatment effect, as this measure appears to reduce the indirect treatment effect by about nine percent.

Discussion

This study describes an initiative that is low-cost and relatively easy to implement at scale while still demonstrating promise for improving teacher preparation. The core of the initiative involved the use of administrative data to identify the most instructionally effective and experienced teachers in districts and then to share recommendation lists that encourage district leaders to target these teachers in their recruitment of CTs. Leaders in districts that were randomly assigned to use the recommendation lists were able to recruit substantially more effective and experienced CTs (by 0.4-0.7 standard deviation units, depending upon the outcome

and model). Policymakers in Tennessee, more than most other states, already prioritize recruiting instructionally effective CTs, as evidenced by the fact that they are one of only a few states that set minimum requirements for evaluation scores in order for teachers to serve as CTs. In the context of this state policy, the success of our initiative in raising the average effectiveness and experience of the pool of CTs by a marked degree demonstrates the potentially widespread applicability of this strategy.

Taking a skeptical perspective, one might view this study's first set of findings to be unexceptional; it may not seem groundbreaking that recruiters are able to recruit more instructionally effective and experienced teachers when told which are most effective and experienced! When we began this initiative, however, there was uncertainty among state, district, and TTU leaders about whether they had already tapped the local supply of available, instructionally effective teachers in needed subjects, counties, and grade levels to serve. After all, district leaders already had access to evaluation data on teachers and were already prompted to target instructionally effective teachers as per state policy. Ronfeldt, Goldhaber, et al. (2018) showed that, even without recommendation lists, program and district/school leaders across Tennessee were already recruiting CTs that were meaningfully more effective and experienced than other teachers in the state. In other words, recruiters were already doing quite well, but could they do better? Substantial doubts were also raised by stakeholders regarding evidence that instructionally effective teachers might be unwilling or unable to serve (Mullman & Ronfeldt, 2019), at least in part because of local concerns that serving as CTs might harm evaluation scores (Goldhaber et al., 2018b; Ronfeldt, et al., 2020; SAS Institute, 2014).

Results from our initiative suggest, then, that the above premise was not true: there were more instructionally effective and experienced teachers available and willing to serve as CTs in

needed districts/subjects/grades. Given that recruiters were able to enlist them as part of the initiative without offering any additional incentives to serve, a reasonable conclusion is that the most instructionally effective and experienced teachers were not already being asked to serve. This raises another set of questions, though, that need to be investigated in the future: why weren't the most instructionally effective and experienced teachers already being asked to serve? Was it because recruiters did not know who to target? This seems unlikely, given that district leaders have access to the same evaluation and administrative data that we did. Perhaps they had access to the information but did not have a systematic method, like our algorithm, for identifying the most instructionally effective and experienced teachers in needed endorsement areas. Alternatively, it might be the case that district leaders were using data to successfully recruit, but the breakdown was in districts where program leaders, who did not have access to evaluation data, took primary responsibility for recruitment.[13] Another possibility is that all recruiters knew who were the most instructionally effective and experienced teachers but instead used other criteria for their recruitment – for example, reputations about which CTs were the best and most supportive mentors of adult learners, PSTs' familiarity with the CT or school setting, or CTs' existing relationships with TEPs (Mullman & Ronfeldt, 2019). More research is needed to understand why teachers at the top of the recommendation lists were not already being targeted.

Beyond our specific setting, there are a number of possible reasons why other states may not currently prioritize the recruitment of CTs with strong evaluation scores in general. First, a number of states do not collect observation ratings or value-added measures for all teachers in the state, so such a policy would be infeasible. In other contexts, performance information is available, but there may be skepticism about its usefulness or validity, even as findings from this and other recent studies counter such skepticism. Relatedly, there is a common perception among

many teacher educators (and stakeholders generally) that it is more critical to PSTs' learning to have CTs that are supportive mentors of young adults than exceptional teachers of P-12 students. Though we agree in principle with the emphasis on coaching quality, our study adds to growing evidence that recruiters should also prioritize instructionally effective teachers of P-12 students.

The initiative also seemed to benefit PSTs, as those who worked with CTs recruited using recommendation lists felt significantly better prepared to teach at the end of their preparation programs (by 0.5–0.7 standard deviation units, depending upon the outcome and model). This result is notable, as it suggests that deliberately leveraging an evidence-based feature of teacher education – the level of effectiveness and experience of CTs – can have a causal impact on PSTs' sense of preparedness to teach. Given that teacher education programs consist of a web of interacting and interdependent components, one might expect program improvement to require a systemic, rather than a feature-specific, approach to change; however, in this instance, we found this to not be the case. These results provide support for an approach to improving feelings of preparation that targets specific, evidence-based features as levers for change. Whether to implement a systemic or feature-specific approach may depend, though, on the kinds of features being targeted. For example, a shift in content focus (e.g., supporting social-emotional learning) might require a more systemic approach – where fieldwork and coursework experiences are collectively revised to ensure coherence across them.

It is also important to underscore that our focus on CTs' instructional effectiveness was not idiosyncratic but instead empirically grounded. As described in the introduction, at least four studies in three different labor markets have found positive associations between CTs' instructional effectiveness and PSTs' instructional effectiveness or feelings of preparedness to teach (Goldhaber et al., 2018a; Matsko et al., 2020; Ronfeldt, Brockman, & Campbell, 2018;

Ronfeldt, Matsko, et al., 2020). Only one prior study, though, went beyond correlational evidence to use an experimental design to test whether these effects are truly causal. In that working paper, Ronfeldt, Goldhaber, et al. (2018) find evidence that PSTs who were randomly assigned to "high-index" placements – combining instructionally effective CTs with placement schools that have lower teacher turnover and stronger achievement gains – reported better quality and more frequent coaching from their CTs; they also reported feeling somewhat better prepared to teach, though not at statistically significant levels. Our results are somewhat reversed, in that we find few significant effects on coaching activities but significant, positive, and large (about twice the magnitude reported in Ronfeldt, Goldhaber, et al., 2018) effects on PSTs' sense of preparedness to teach. A distinction between these studies, though, is that the recruitment strategy in the earlier study targeted promising school placements alongside promising CTs, while the current initiative targeted promising CTs exclusively.

The present study is the first then, to our knowledge, to provide causal evidence that recruiting more effective and experienced CTs improves PSTs' self-perceived preparedness to teach. That said, in our view, helping PSTs feel better prepared is not enough, as it does not always predict becoming instructionally effective once in the classroom (Ronfeldt, Matsko, et al., 2020). For example, it is possible that, when more instructionally effective teachers serve as CTs, they tend to maintain more control over setting and maintaining classroom procedures. This would be consistent with our finding that PSTs in districts randomly assigned to receive the recommendation lists reported somewhat less autonomy over instructional decisions. As a result of CTs maintaining control, classrooms may run more smoothly, thus causing PSTs to feel more prepared to set and maintain classroom procedures but not necessarily be any more effective at doing so on their own. In future work, we will examine whether or not graduates who completed

their clinical training in treatment districts are also more instructionally effective (based upon state evaluation measures) during the first year of teaching.

Finally, while PSTs seemed to benefit, on average, from being assigned more instructionally effective and experienced CTs, we are less clear on how – whether (1) through modeling more effective teaching, (2) through better coaching practices, where more effective teachers are able to translate their teaching skills with P-12 students into stronger coaching skills with learning teachers (PSTs), or (3) through both mechanisms. When we examined whether or not PSTs in treatment districts reported better or more frequent coaching, results were mixed. We found little evidence that better coaching practices were associated with or mediated the relationship we observed between treatment assignment and PST feelings of preparedness; the coaching practices we measured reduced the effect of treatment on PSTs' feelings of preparedness by a small percentage, but the main effects were still large and significant with the inclusion of the coaching measures. Moreover, when we included indirect paths for modeling and coaching, our results suggested that the former reduced more of the effect of treatment on outcomes than the latter. Our measure of instructional modeling consistently reduced about twenty-five percent of the total treatment effect, while the measure of instructional coaching only reduced up to twelve percent of the treatment effect when using PSTs' reports of coaching frequency as a proxy for instructional coaching. In other words, we found less evidence in support for explanation (2) – that recruiting more instructionally effective CTs impacts PSTs' preparedness to teach through improving the coaching that PSTs receive from their CTs.

One might be tempted to conclude then that (1) must be true, but reaching such a conclusion would also, we believe, be premature. First, while the coaching measures we use do not seem to explain much of the main effect of treatment on PSTs' preparedness, it is possible

that we do not observe other kinds of coaching that could explain these relationships. Future research might consider investigating other forms of coaching and using different coaching measures, including measures of observed, rather than survey-reported, coaching practice. Second, while we have evaluation measures of CTs' instructional effectiveness, we do not have measures of whether PSTs are vicariously learning from the instruction that their CTs are modeling or demonstrating. It might be that some PSTs do not actually observe their CTs' instruction or that they observe but do not attend to or learn from the aspects of instruction that they perhaps should. Even if we did have adequate measures for PSTs' vicarious learning from CTs' instructional practice, we would need to conduct mediation analyses with PST vicarious learning measures in order to determine the degree to which these may explain the main effects of treatment on PSTs' preparedness to teach. More work is needed to identify the mechanism by which being assigned more instructionally effective teachers causes PSTs to feel more prepared to teach. One possibility would involve the collection of observational data on the coaching practices (e.g., feedback during coaching conferences, frequency of modeling and co-teaching) of more versus less instructionally effective teachers (of P-12 students) to examine whether the kinds and quality of coaching are qualitatively different. Finally, the results of these mediation analyses move away from the causal inference framework that allows us to interpret our other results as causal estimates of receiving recommendation lists. These mediation analyses are exploratory in nature, and our experiment was not designed to estimate causal relationships between coaching and modeling mediators and PSTs' feelings of preparedness. Future work should consider experimental designs that produce causal estimates of the effects of coaching and modeling on PSTs' outcomes.

Even though the mediation models suggest that the treatment effects on feelings of preparedness might flow more through instructional modeling than instructional coaching, it is important to highlight some limitations of these analyses. First, if our measure of coaching is less reliable than our measure of modeling, then the difference in the mediation magnitude could disappear with a more reliable measure of CTs' coaching. Similarly, about half of the treatment effect is still explained by the direct relationship between the treatment indicator and feelings of preparedness. Effectively, these results suggest that about half of our treatment effect remains unexplained by our measures of instructional modeling or instructional coaching. This could be explained in several ways: (1) our factor scores could not fully capture the extent and intensity of CTs' instructional coaching, (2) our recruitment index only captures a fraction of CTs' instructional modeling, or (3) some other channel through which treatment (i.e., the selection of CTs by district) might impact PSTs' feelings of preparedness that we have not measured. Unfortunately, we are unable to test these hypotheses in the current study. We believe that this could be a fruitful area of focus for future research.

While it would be useful to better comprehend the mechanism (modeling, coaching, or some combination of the two), the main contributions of this present study are (1) to offer the first evidence, to date, that more instructionally effective and experienced CTs, in fact, have a causal impact on PSTs' preparedness to teach and (2) to present a feasible, low-cost method for raising the average effectiveness and experience of the pool of CTs simply by providing leaders with actionable information in the form of strategic recommendation lists. Consistent with prior correlational analyses, this study supports existing policies and practices, like those in the state of Tennessee, that set minimum requirements for how instructionally effective teachers must be in order to serve as CTs. Building on support for these minimum requirement policies, this study

presents evidence that providing improved information can induce changes in the pool of CTs over and above the minimum requirements to the direct benefit of the PSTs during their clinical experiences.

Endnotes

[1] The term "model" and "modeling" have been used to represent many different activities. We here refer to the CT as a "model" in a very simplistic and rudimentary sense – where, through enacting teaching aimed at P-12 students, the CT demonstrates teaching to the PST, regardless of the degree to which this demonstration is deliberately meant to teach anything specifically to the PST. Others have written about forms of modeling where CTs deliberately structure their enactments of teaching in ways that are meant to demonstrate very specific aspects of practice and where the enactments are structured in a way to ensure that the PST observes and learns from these enactments; we are not here referring to these more deliberate forms of modeling.

[2] PSTs in TTU completed their residency experiences in only one main clinical placement, except for 20 PSTs who had musical or special education placements and needed to complete a second clinical placement to fulfill their specific credentialing requirements. We generated new recommendation lists for these PSTs. TTU followed the same recruitment procedures for this new cohort of mentors as the first recruitment drive.

[3] TTU uses the term "mentor" instead of "cooperating teacher (CT)" and "resident" instead of "preservice student teacher (PST)." We use "CT" and "PST" because these are more common terms in the teacher education literature and in order to be consistent with the terminology used in the rest of this manuscript.

[4] Because we had to generate these lists many months prior to the beginning of the academic year, centralized information on which teachers were assigned to teach which subjects, courses, and grade levels were not yet available. Thus, we used TDOE course files from prior years to identify all the subjects, courses, and grade levels that teachers had previously been assigned to predict which teachers might be potential matches for the relevant blocks/PSTs. Because teachers sometimes switch subjects and grades from one year to the next, recommendation lists sometimes included teachers who did not actually match the needed subject-grade blocks. In these cases, administrators in charge of recruitment were advised to note such misclassifications and then move to the next teachers in the recommendation lists.

[5] In Tennessee, most teachers are evaluated using the Tennessee Educator Acceleration Model (TEAM) rubric. This rubric evaluates teaching practice along four domains (i.e., Planning, Instruction, Environment, and Professionalism) on a 1- to 5-point scale (from "Significantly Below Expectations" to "Significantly Above Expectations"). All teachers in the state are evaluated at least once each school year. About 20% of teachers in the state are evaluated using different observation rubric than the TEAM. We rely on the equating work done at the Tennessee Department of Education when using observation scores from districts that use different observation rubrics.

[6] Tennessee uses the Tennessee Value-Added Assessment System (TVAAS) to calculate teachers' contributions to test scores. The models used to calculate teachers' VAM scores differ from traditional econometric models insofar that they do not directly include student demographic characteristics in the regression models. Instead, student growth scores are calculated using lagged growth models at the teacher level. More technical information on the modeling of TVAAS scores is available here: https://tvaas.sas.com/

[7] We use recruitment fields as a proxy for endorsement area for teachers. We identify recruitment fields using teacher assignments at the course level, and we infer which endorsement teachers are likely to have. We have cross-referenced the crosswalk between courses and endorsements with

our TTU partners to ensure that recommended recruitments would fulfill the requirements for being recommended for a specific endorsement.

[8] We chose these weights for the measures for two reasons. We wanted to use observation ratings and teacher value added measures because these two measures are the only measures of teacher effectiveness that are available state-wide for most teachers. There is also substantial prior literature indicating that teachers improve with experience, particularly early in their careers (e.g., Papay & Kraft, 2015); thus, we thought it important to also include experience as part of our composite measure. Moreover, Tennessee law has prohibited sharing of teacher evaluation data in most cases, so adding experience also ensured our composite measure was not solely a measure of instructional effectiveness. The inclusion of teacher experience made our recruitment index different enough from teacher evaluation data according to the state department's legal department. This made it possible for us to share our recruitment data with Tennessee Technological University for the purpose of recruiting CTs. We have run several robustness checks to test the extent to which our preferred recruitment index weights would influence the overall ranking of teachers. All these reweighted teacher rankings are highly correlated with our preferred weighting scheme.

[9] We also tested alternative specifications that included district-level random effects: a three-level nested structure with field-level random effects, a crossed random effects structure with field-level random effects, or a two-level structure that nested PSTs within districts. We consistently found that the district-level random effects did not explain enough variation in our outcomes of interest to justify their inclusion in our models.

[10] We tested whether there were significant differences between estimates from our preferred models (i.e., recruitment field fixed effects) and this specification (i.e., recruitment field random effects) using Hausman tests. All tests failed to reject the null hypothesis that there is no systematic difference between the two estimators.

[11] The interpretation of the results for models with recruitment field fixed-effects should be interpreted as the within-field effects of receiving recommendation lists on overall CT quality. These models account for possible unobserved differences in recruiting strategies among recruitment fields. For example, it might be easier to recruit a CT for an elementary education placement than agriculture education one given the larger number of possible CTs for elementary education placements.

[12] Given its primary reliance on evaluation metrics, which can be noisy measures from year to year that are susceptible to regression to the mean, we explored whether shifting the three-year window during which we construct the recruitment index resulted in substantively different estimates of this overall contrast. Moving the window forward one (2016-2018) and two years (2017-2019) saw a modest reduction in the magnitude of this contrast (between 15-20%), but all estimates remained qualitatively similar and statistically significant at the $p < 0.01$ level.

[13] The program took primary responsibility for recruitment in six of the twenty-five districts participating in this initiative (two in treatment and four in control). We use a difference-in-differences approach to compare the treatment contrast between recruitment strategies. We find the treatment contrast for placements where program leaders (rather than district office leaders) to actually be smaller in magnitude. These results contradict the hypothesis that the contrast would be greater in districts where program leaders take primary responsibility for recruitment due to the fact that – prior to the study – they did not have access to evaluation data so were less able to select CTs based upon measures of instructional effectiveness (whereas district leaders

had access to evaluation data). These results, though, cannot be interpreted as causal effects because our randomization strategy was not designed to stratify treatment within recruitment strategies.

References

Bastian, K. C., Patterson, K. M., & Pan, Y. (2018). Evaluating teacher preparation programs with teacher evaluation ratings: Implications for program accountability and improvement. *Journal of Teacher Education*, *69*(5), 429–447. https://doi.org/10.1177/0022487117718182

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Glenn, W. J. (2006). Model versus mentor: Defining the necessary qualities of the effective cooperating teacher. *Teacher Education Quarterly*, *33*(1), 85–95.

Goldhaber, D., Krieg, J., & Theobald, R. (2018a). *Effective like me? Does having a more productive mentor improve the productivity of mentees?* (Calder Working Paper No. 208-1118-1). Retrieved from National Center for Analysis of Longitudinal Data in Educational Research website: https://caldercenter.org/sites/default/files/CALDER%20WP%20208-1118-1.pdf

Goldhaber, D., Krieg, J., & Theobald, R. (2018b). *The costs of mentorship? Exploring student teaching placements and their impact on student achievement* (CEDR Working Paper No. 03132018-1-1). Retrieved from Center for Education Data & Research website: http://www.cedr.us/papers/working/CEDR%20WP%202018-1.pdf

Gordon, M. F., Jiang, J. Y., Kapadia Matsko, K., Ronfeldt, M., Greene Nolan, H. G., & Reininger, M. (2018). *On the path to becoming a teacher: The landscape of student teaching in Chicago Public Schools*. UChicago Consortium on School Research. https://consortium.uchicago.edu/sites/default/files/2018-10/On%20the%20Path%20to-Aug2018-Consortium.pdf

Grossman, P., Ronfeldt, M., & Cohen, J. J. (2012). The power of setting: The role of field experience in learning to teach. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook, Vol 3: Application to learning and teaching* (pp. 311–334). American Psychological Association. https://doi.org/10.1037/13275-023

Hoy, W. K., & Woolfolk, A. E. (1990). Socialization of student teachers. *American Educational Research Journal*, *27*(2), 279–300. https://doi.org/10.3102/00028312027002279

Kang, H. (2020). The role of mentor teacher–mediated experiences for preservice teachers. *Journal of Teacher Education*. https://doi.org/10.1177/0022487120930663

Koerner, M., Rust, F. O., & Baumgartner, F. (2002). Exploring roles in student teaching placements. *Teacher Education Quarterly*, *29*(2), 35–58.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Krieg, J. M., Goldhaber, D., & Theobald, R. (2020). Teacher candidate apprenticeships: Assessing the who and where of student teaching. *Journal of Teacher Education*, *71*(2): 218-232. https://doi.org/10.1177/0022487119858983

Krieg, J. M., Theobald, R., & Goldhaber, D. (2016). A foot in the door: Exploring the role of student teaching assignments in teachers' initial job placements. *Educational Evaluation and Policy Analysis*, *38*(2), 364–388.

Maier, A., & Youngs, P. (2009). Teacher preparation programs—and teacher labor markets: How social capital may help explain teachers' career choices. *Journal of Teacher Education*, *60*(4), 393–407. https://doi.org/10.1177/0022487109341149

Matsko, K. K., Ronfeldt, M., Nolan Greene, 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).

Mullman, H. & Ronfeldt, M. (2019). Walking around the puddles: The landscape of clinical preparation in Tennessee. Unpublished manuscript.

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *J. Public Econ., 130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Rickenbrode, R., Drake, G., Pomerance, L., & Walsh, K. (2018). *2018 Teacher Prep Review*. National Council on Teacher Quality.

Ronfeldt, M. (2012). Where should student teachers learn to teach? *Educational Evaluation and Policy Analysis, 34*(1), 3–26. https://doi.org/10.3102/0162373711420865

Ronfeldt, M., Bardelli, E., Brockman, S. L., & Mullman, H. (2020). Will mentoring a student teacher harm my evaluation scores? Effects of serving as a cooperating teacher on evaluation metrics. *American Educational Research Journal*, *57*(3), 1392–1437. https://doi.org/10.3102/0002831219872952

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., Goldhaber, D., Cowan, J., Bardelli, E., Johnson, J., & Tien, C. D. (2018). *Identifying promising clinical placements using administrative data: Preliminary results from ISTI placement initiative pilot* (CALDER Working Paper 189). Retrieved from National Center for Analysis of Longitudinal Data in Educational Research website: https://caldercenter.org/publications/identifying-promising-clinical-placements-using-administrative-data-preliminary-results

Ronfeldt, M., Matsko, K. K., Greene Nolan, H., & Reininger, M. (2020). Three different measures of graduates' instructional readiness and the features of preservice preparation that predict them, *Journal of Teacher Education*. Advance online publication. https://doi.org/10.1177/0022487120919753

Rozelle, J. J., & Wilson, S. M. (2012). Opening the black box of field experiences: How cooperating teachers' beliefs and practices shape student teachers' beliefs and practices. *Teaching and Teacher Education*, *28*(8), 1196–1205. https://doi.org/10.1016/j.tate.2012.07.008

SAS Institute. (2014). *Preliminary report: The impact of candidates on teacher value-added reporting*. Cary, NC.

Schwille, S. A. (2008). The professional practice of mentoring. *American Journal of Education*, *115*(1), 139–167. https://doi.org/10.1086/590678

St John, E., Goldhaber, D., Krieg, J., & Theobald, R. (2018). *How the match gets made: Exploring student teacher placements across teacher education programs, districts, and schools* (CEDR Working Paper No. 10052018-1-1). Retrieved from Center for Education Data & Research website:

http://t.cedr.us/papers/working/CEDR%20Working%20Paper%20No.%2010052018-1-1.pdf

Yendol-Hoppey, D. (2007). Mentor teachers' work with prospective teachers in a newly formed professional development school: Two illustrations. *Teachers College Record*, *109*(3), 669–698.

Zeichner, K., & Gore, J. (1990). Teacher socialization. In W. R. Houston (Ed.), *Handbook of research on teacher education* (pp. 329–348). McMillan.

Table 2: Balance Check on District Characteristics

Variable	All	Control	Treatment	Diff	Effect Size
Panel A: Student Characteristics					
% African American	6.168	5.758	6.578	0.820	0.101
% Hispanic	6.621	6.852	6.389	-0.463	0.084
% Asian	0.995	1.267	0.722	-0.545	0.546
% Native American	0.296	0.296	0.296	0.000	0.001
% White	85.783	85.651	85.915	0.263	0.021
% Hawaiian or Pacific Islander	0.116	0.133	0.100	-0.033	0.409
% Free or Reduced-Price Lunch	36.413	34.577	38.249	3.673	0.419
% Students with Disabilities	15.047	14.912	15.182	0.270	0.155
% English Language Learners	0.524	0.571	0.476	-0.095	0.177
Panel B: Potential CT Evaluation Scores					
Mean Observation Rating	3.987	3.993	3.980	-0.013	0.060
Mean Instructional Domain Rating	3.877	3.876	3.878	0.003	0.015
Mean Environment Domain Rating	4.353	4.348	4.358	0.010	0.041
Mean Planning Domain Rating	3.937	3.985	3.896	-0.089	0.301
Mean Professionalism Rating	4.241	4.216	4.263	0.047	0.184
Mean VAM	0.031	0.059	0.003	-0.057	0.690+
90th Percentile Observation Rating	4.614	4.604	4.623	0.019	0.082
90th Percentile Instructional Domain Rating	4.473	4.490	4.458	-0.031	0.120
90th Percentile Environment Domain Rating	4.940	4.925	4.955	0.030	0.206
90th Percentile Planning Domain Rating	4.640	4.692	4.597	-0.094	0.229
90th Percentile Professionalism Rating	4.917	4.900	4.932	0.032	0.182
90th Percentile VAM	0.333	0.352	0.314	-0.038	0.287
N	*24*	*12*	*12*		

Note. This table reports the results of a balance check on district characteristics at time of randomization. School-level student characteristics are calculated using publicly available school data from Tennessee. Teacher evaluation data includes tests for the average evaluation scores of teachers in the same district and subject as requested placements as well as the 90th percentile of each score distribution. The 90th percentile tests for the availability of highly effective CTs in treatment and control districts. A joint test of significance across the five covariates is non-significant for both panels (Panel A: $\chi^2(9) = 8.771$, $p = 0.459$; Panel B: $\chi^2(12) = 8.463$, $p = 0.748$). + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1. Cooperating Teacher (CT) Characteristics

Variable	(1) Non-Initiative Teachers	(2) All Cooperating Teachers	(3) Treatment Cooperating Teachers	(4) Control Cooperating Teachers
2015 Observation Ratings	3.852	4.266	4.336	4.183
	(.580)	(.491)	(.451)	(.526)
	7179	*147*	*80*	*67*
2016 Observation Ratings	3.893	4.298	4.406	4.175
	(.581)	(.439)	(.379)	(.473)
	8444	*159*	*85*	*74*
2017 Observation Ratings	3.945	4.342	4.427	4.247
	(.580)	(.429)	(.401)	(.442)
	9168	*159*	*84*	*75*
2015 TVAAS Scores	-.043	.206	.309	.061
	(.373)	(.316)	(.317)	(.257)
	3510	*65*	*38*	*27*
2016 TVAAS Scores	-.015	.265	.265	.264
	(.331)	(.410)	(.431)	(.393)
	755	*26*	*17*	*9*
2017 TVAAS Scores	-.004	.251	.313	.168
	(.268)	(.230)	(.217)	(.227)
	2728	*51*	*29*	*22*
2017 Years of Experience	10.986	15.656	17.941	13.067
	(8.991)	(9.858)	(10.330)	(8.655)
	9376	*160*	*85*	*75*
Observation Scores Index	-.162	.515	.675	.334
	(.982)	(.719)	(.648)	(.756)
	9259	*160*	*85*	*75*
TVAAS Index	-.034	.228	.320	.110
	(.284)	(.296)	(.298)	(.251)
	4352	*77*	*43*	*34*
Experience Index	-.035	.449	.704	.160
	(.988)	(1.086)	(1.144)	(.943)
	9376	*160*	*85*	*75*
Recruitment Index	-.113	.531	.731	.304
	(.742)	(.629)	(.548)	(.641)
	9310	*160*	*85*	*75*

Note. This table reports summary statistics for the CT recruitment index. Column 1 reports the estimates for the potential CTs who taught requested field recruitments in districts participating in the initiative. Columns 2 through 4 report the estimates for the initiative's CTs. Standard deviations are in parentheses. Cell counts are in italics.

Table 3. Contrast of CT Recruitment Index Measures between Treated and Control Districts

Contrast	**Observation Ratings**				
	Average	**Instruction**	**Environment**	**Planning**	**Prof.**
Contrast	0.184	0.103	0.057	0.004	0.198**
	(0.115)	(0.102)	(0.080)	(0.151)	(0.069)
Adj. Contrast	0.231+	0.159	0.075	0.040	0.235*
	(0.130)	(0.119)	(0.093)	(0.167)	(0.085)
Std. Contrast	0.332	0.184	0.109	0.012	0.336*
	(0.207)	(0.190)	(0.138)	(0.244)	(0.121)
Adj. Std. Contrast	0.415+	0.288	0.126	0.062	0.397*
	(0.234)	(0.218)	(0.158)	(0.258)	(0.141)
	VAM				
	Average	**Mathematics**	**English Language Arts**		**Experience**
Contrast	0.208**	0.411**	0.222**		5.007**
	(0.065)	(0.136)	(0.077)		(1.498)
Adj. Contrast	0.215**	0.424**	0.199*		5.081**
	(0.072)	(0.140)	(0.080)		(1.595)
Std. Contrast	0.654**	0.967**	0.744**		0.558**
	(0.203)	(0.311)	(0.257)		(0.178)
Adj. Std. Contrast	0.683**	0.979**	0.702*		0.570**
	(0.229)	(0.318)	(0.268)		(0.178)

Note. This table reports the contrast between treatment and control CTs on evaluation scores. Adjusted estimates include fixed effects for PSTs' recruitment field requests. Standardized scores are calculated at the state level within recruitment field requests. Clustered standard errors at the block level in parentheses. + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 4. PST Post-Survey Differences between Treatment and Control Districts

Survey Factor	**(1) Preferred Model**	**(2) Bootstrap S.E.**	**(3) Pre-Survey Control**	**(4) R.E. Model**
Feeling of Preparedness – Teaching Skills	0.593*	0.593*	0.451*	0.579***
	(0.226)	(0.250)	(0.207)	(0.155)
Preparedness in Questioning Skills	0.637*	0.637*	0.497*	0.626***
	(0.230)	(0.291)	(0.207)	(0.161)
Preparedness in Other Instructional Skills	0.548*	0.548*	0.406+	0.531***
	(0.225)	(0.256)	(0.214)	(0.151)

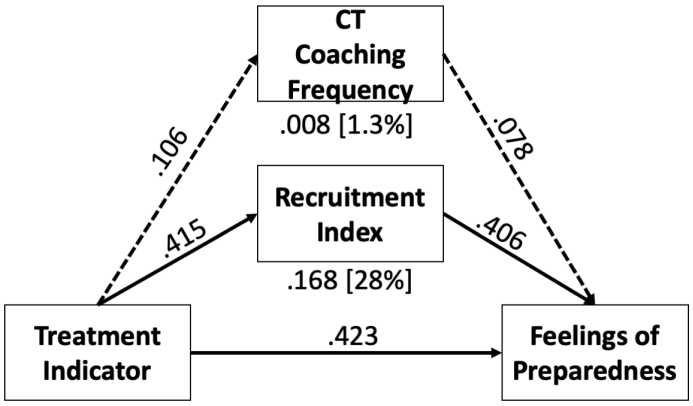
Notes. + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 5. Differences in Coaching between Treatment and Control Districts

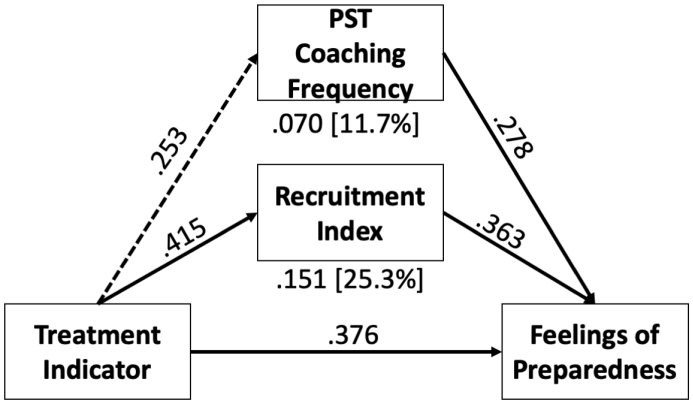
Survey Factor	(1) Preferred Model	(2) Bootstrap S.E.	(3) Pre-Survey Control	(4) R.E. Model
PST Surveys				
Frequency of Coaching Practices	0.181	0.181	0.182	0.181*
	(0.147)	(0.166)	(0.201)	(0.083)
Common Coaching Practices	0.143	0.143	0.151	0.143
	(0.184)	(0.195)	(0.231)	(0.102)
Data-Driven Coaching Practices	0.236	0.236	0.282	0.240***
	(0.201)	(0.216)	(0.267)	(0.067)
Collaborative Coaching Practices	0.205+	0.205	0.152	0.174
	(0.111)	(0.132)	(0.153)	(0.114)
Modeling Coaching Practices	0.141	0.141	0.144	0.179+
	(0.186)	(0.219)	(0.260)	(0.093)
Coaching Satisfaction	-0.143	-0.143	-0.158	-0.105
	(0.171)	(0.211)	(0.269)	(0.066)
Support and Feedback	-0.181	-0.181	-0.178	-0.146*
	(0.170)	(0.178)	(0.267)	(0.066)
Autonomy and Encouragement	-0.105	-0.105	-0.138	-0.064
	(0.179)	(0.198)	(0.277)	(0.070)
CT Surveys				
Frequency of Coaching Practices	0.241	0.241	0.241	0.207**
	(0.193)	(0.245)	(0.301)	(0.077)
Debriefing	-0.037	-0.037	-0.037	-0.103
	(0.203)	(0.152)	(0.173)	(0.153)
Developing Practice	0.277	0.277	0.277	0.282+
	(0.196)	(0.209)	(0.229)	(0.145)
Collaborative Coaching Practices	0.062	0.062	0.062	0.053
	(0.196)	(0.230)	(0.251)	(0.087)
Coaching Frequency in Instruction Domain	0.236	0.236	0.236	0.153
	(0.202)	(0.157)	(0.201)	(0.133)

Notes. CT outcomes include controls for administration cohort. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

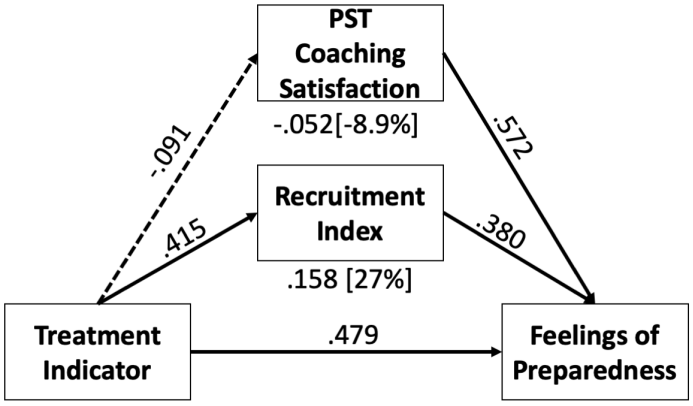
Figure 1: Treatment Mediation through CT Modeling and Coaching
Panel A. Mediation through Coaching Frequency—CT Surveys



Panel B. Mediation through Coaching Frequency—PST Surveys



Panel C. Mediation through Coaching Satisfaction—PST Surveys



Notes. Models inlcude controls for CT survey administration cohort and recruitment field. Solid arrows indicate significant paths at the 0.05 level; short dashed lines indicate paths significant at the 0.10 level; long dashed lines indicate non-significant paths.

Appendices

Appendix Table 1. Preservice Student Teacher (PST) Characteristics

Variable	All	Control	Treatment	Diff	Effect Size
Female	0.834	0.853	0.816	-0.036	0.097
N	*193*	*95*	*98*		
White	0.974	0.979	0.969	-0.010	0.060
N	*193*	*95*	*98*		
Current GPA	3.456	3.447	3.464	0.017	0.051
N	*193*	*95*	*98*		
ACT Score	22.804	22.581	23.033	0.452	0.149
N	*184*	*93*	*91*		
Praxis Score	168.944	169.122	168.775	-0.347	0.032
N	*173*	*84*	*89*		

Note. This table reports the results of a balance check on PST characteristics at time of randomization. A joint test of significance across the five covariates is non-significant ($\chi^2(5) = 2.352$, $p = 0.779$). + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 2. Survey Non-Response Rates by Treatment Condition

Variable	Missing Pre-Survey	Missing Post-Survey	Missing CT Survey
Treatment	0.001	-0.074	-0.078
	(0.067)	(0.061)	(0.049)
Constant	0.331^{***}	0.584^{***}	0.290^{***}
	(0.030)	(0.038)	(0.032)
N	*172*	*172*	*172*
R^2	0.120	0.134	0.152
Adjusted R^2	0.041	0.057	0.076

Note. This table reports the difference in survey non-response rates between control and treatment conditions. We calculate these missing rates by regressing an indicator for survey non-response on the treatment indicator. We include recruitment field fixed effects and cluster standard errors at the school district level for consistency with our preferred estimation models. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix Table 3. Pre-Survey Differences between Treatment and Control Districts

Survey Factor	(1) Preferred Model	(2) Bootstrap S.E.	(3) R.E. Model
Feeling of Preparedness - Teaching Skills	0.345	0.345	0.290*
	(0.282)	(0.317)	(0.117)
Preparedness in Questioning Skills	0.356	0.356	0.328**
	(0.295)	(0.324)	(0.114)
Preparedness in Other Instructional Skills	0.333	0.333	0.252*
	(0.273)	(0.284)	(0.125)

Note. + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

Appendix Table 4. Treatment Effect Mediation Through Measures of Coaching and Modeling

Survey Factor	Point Estimate	95% Bootstrap Confidence Interval	Percent Reduction
Panel A: Coaching Frequency Reported by CT			
Direct Treatment Effect	0.423	[0.003, 0.990]	0.706
Indirect Treatment via Coaching Frequency	0.008	[-0.071, 0.026]	0.013
Indirect Treatment via Recruitment Index	0.168	[0.041, 0.341]	0.280
Total of Indirect Effects	0.177	[0.046, 0.448]	0.295
Panel B: Coaching Frequency Reported by PST			
Direct Treatment Effect	0.376	[0.081, 0.602]	0.630
Indirect Treatment via Coaching Frequency	0.070	[-0.105, 0.240]	0.117
Indirect Treatment via Recruitment Index	0.151	[0.035, 0.305]	0.253
Total of Indirect Effects	0.221	[0.050, 0.462]	0.370
Panel C: Coaching Satisfaction Reported by PST			
Direct Treatment Effect	0.479	[0.376, 0.553]	0.819
Indirect Treatment via Satisfaction with Coaching	-0.052	[-0.341, 0.090]	-0.089
Indirect Treatment via Recruitment Index	0.158	[0.053, 0.253]	0.270
Total of Indirect Effects	0.105	[-0.140, 0.281]	0.179

Note. This table reports the results of three different structural equation models that estimate the direct and indirect treatment effects (i.e., mediation analyses) through measures of coaching and modeling. Models were estimated in Mplus 7.4 using latent variables for each survey-based measure. We used a WLSMV estimator to accommodate categorical items for the PSTs' feelings of preparedness. All models had good fit indices. All point estimates and confidence intervals are obtained from 100 bootstrapped replications.

Online Appendix A – Psychometric Properties of Our Survey Instruments

We use confirmatory factor analyses to calculate the factor scores for our outcomes of interest. We calculate the factor scores in Stata using the "sem" command. This decision relies on two main assumptions: (1) all observed indicators are continuous variables and (2) all observed indicators are normally distributed. Both these assumptions are somewhat standard for traditional principal component factor analyses but could lead to biased results within an SEM framework (see, Bollen, 1989, for an in-depth treatment of the validity threats in violating these assumptions). Practically, the chi squared fit statistics – and all its derivative fit indices – are sensitive to the violation of the assumption that observed variables are normally distributed. Satorra and Bentler (1994) describe a correction for these fit indices that is robust in small samples and for non-normal data.

We follow a data-driven approach to decide when and if to include error covariance terms in our models when model modification indices suggest that the inclusion of these terms would improve overall model fit. Following modification indices to improve model fit is a double-edged sword. On one hand, the inclusion of error covariance terms allows for the explicit modeling of unobserved factors that could influence participant responses to two questions that are unrelated to the latent factor of interest. On the other hand, these error covariance terms are likely to be sample specific, which might lead to overfitting of the measurement model to our data. We try to address these concerns in two ways. First, we estimate the measurement model parameters using responses from multiple TEPs in the state, some of which did not participate in the Mentors Matter Recruitment initiative. This reduces the risk to overfit our measurement models to specific features of teacher preparation of one specific program. For example, if the methods course in our partner TEP focused on the use of computers in differentiating instruction, we might observe its effects as an error covariance term between questions about preparedness in using computers and differentiating instruction. Using data from multiple TEPs reduces this risk because the effects of this specific focus would "wash out" with the inclusion of responses from other TEPs. Second, we include modification indices only when we can theoretically justify their inclusion in the model. This prevents us from blindly follow the suggestions of our statistical software and to leverage our expertise to improve the measurement models.

Appendix Table A1 reports the fit indices for each factor model that we fit. We include each model's chi squared statistic, root mean squared error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), standardized root mean square residual (SRMR), and the coefficient of determination (CD). Fit indices allow to assess the extent to which the proposed factor model fits the observed participant response data.

Broadly speaking, fit indices fall within two categories, absolute fit and comparative fit indices. Absolute fit indices (e.g., RMESA and SRMR) assume an ideal model with perfect fit and measure the departure of our proposed model from the ideal model. The point estimates for these fit indices are usually expected to be less than an accepted cut-off value for acceptable fit (i.e., RMSEA < 0.06 and SRMS < 0.05). Intuitively, these fit indices assess the average residual size for each survey item. In the case of a CFA, residual terms include the effect of unobserved variables as well as the effects of any stochastic error term on participants' responses.

Comparative fit indices compare our proposed model to a saturated model that includes the maximum number of factors given the data. These fit indices are usually expected to be greater than an accepted cut-off value for acceptable fit (i.e., CFI/TLI > 0.95). Intuitively, these fit indices are a measure of closeness between the saturated model and the proposed model,

where the saturated model is a theoretical model that capture all the variance in the data without any attempt to synthesize information from the data.

Overall, we observe that most of our factors have excellent fit indices with the exception of post-placement coaching frequency, post-placement coaching quality, and CT frequency of coaching on specific subskills. The deviations from excellent fit indices are, however, small enough to signal the measurement models for these factors still have adequate fit to the data. Specifically, high RMSEA values might indicate that responses on some items might be influenced by unobserved covariates beyond the latent construct of interest. We have revised these instruments to address these concerns for future data collections.

Appendix Table A2 reports the reliability estimates for each of the factors and the survey items that load on each factor. Our reliability estimates include two coefficients: the alpha coefficient and the intra-item correlation coefficient. As a rule of thumb, a measure for which alpha is greater than 0.70 and IIC is between 0.15 and 0.50 is said to have acceptable observed reliability. Overall, most of our factors have acceptable to excellent reliability estimates.

Online Appendix Table A1 – Fit Indices for Each Measurement Model

Fit Index		Pre-Prep. Subskills	Post-CT Freq.	Post-Satisfied with CT	Post-Prep. Subskills	CT – Freq.	CT – Freq. Subskills
Chi Squared	Value	48.944	62.492	77.798	51.061	50.127	65.023
	df	42	45	61	43	40	41
	p	0.214	0.043	0.072	0.186	0.131	0.010
RMSEA	Lower Bound	0.026	0.052	0.044	0.037	0.036	0.068
	Estimate	0.028	0.052	0.082	0.053	0.020	0.082
	Upper Bound	0.075	0.103	0.123	0.108	0.077	0.135
	CFI	0.993	0.983	0.988	0.989	0.985	0.970
	TLI	0.991	0.976	0.984	0.985	0.980	0.959
	SRMR	0.038	0.056	0.061	0.038	0.050	0.069
	CD	0.962	1.000	0.994	0.975	0.962	0.931

Online Appendix Table A2 – Survey Measures Reliability Estimates and Factor Loadings

Item	Loading	S.E.
Panel A - PST Pre-Survey - Feeling of Preparedness - Teaching Skills		
Sub-Factor: Preparedness in Questioning Skills (α = 0.879, IIC = 0.592)		
b. Plan question sequences that help students develop deep conceptual understanding	0.683***	(0.046)
e. Ask questions that require students to discuss and/or write out their developing thoughts	0.797***	(0.031)
i. Develop questions that prompt students to grapple with the elements most necessary for understanding a text or concept	0.793***	(0.030)
j. Challenge students to wrestle with deep questions by providing adequate wait time	0.781***	(0.029)
k. Challenge all students by using strategies for calling on all students equitably	0.786***	(0.032)
Sub-Factor: Preparedness in Other Instruction Skills (α = 0.873, IIC = 0.533)		
a. Focus on essential information when presenting content	0.670***	(0.041)
c. Provide activities and materials that are relevant to students' lives	0.780***	(0.034)
d. Provide examples, illustrations, analogies, and labels for new concepts and ideas	0.714***	(0.042)
f. Plan activities that build curiosity	0.732***	(0.039)
g. Present content using visuals that establish the purpose of the lesson	0.810***	(0.037)
h. Incorporate multimedia, technology, and resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.)	0.676***	(0.045)
Covariance Structure		
Questioning with Other	0.907***	(0.020)
Residual for Item a with Residual for Item b	0.321***	(0.063)
Panel B - PST Post-Survey – Frequency of Coaching Practices		
Sub-Factor: Common Coaching Practices (α = 0.867, IIC = 0.765)		
a. Observe you teach	0.866***	(0.029)
c. Prompt you to practice specific aspects of teaching during a lesson	0.868***	(0.031)
Sub-Factor: Data-Driven Coaching Practices (α = 0.934, IIC = 0.702)		
i. Share data or evidence about lessons s/he observed you teach	0.820***	(0.034)
j. Ask you reflective questions	0.840***	(0.029)
k. Analyze student work with you	0.746***	(0.038)
m. Use evaluation data to provide recommendations for improvement	0.855***	(0.024)
n. Provide opportunities outside of regular instruction to practice specific teaching moves	0.772***	(0.047)
l. Share specific next steps for you to work on in order to improve your teaching	0.845***	(0.034)
Sub-Factor: Collaborative Coaching Practices (α = 0.882, IIC = 0.789)		
d. Co-plan a lesson or activity with you	0.999***	(0.034)

e. Co-teach a lesson or part of a lesson with you	0.726***	(0.052)
Sub-Factor: Modeling Coaching Practices (α = 0.791, IIC = 0.654)		
g. Model a specific instructional skill or move when students were not present	0.855***	(0.043)
h. Model a specific instructional skill or move for you during a lesson	0.738***	(0.054)
Covariance Structure		
Common with Data-Driven	0.854***	(0.037)
Common with Collaborative	0.683***	(0.069)
Common with Modeling	0.807***	(0.051)
Data-Driven with Collaborative	0.655***	(0.063)
Data-Driven with Modeling	0.856***	(0.040)
Collaborative with Modeling	0.622***	(0.067)
Residual for Item m with Residual for Item n	0.292***	(0.075)
Residual for Item m with Residual for Item l	0.386***	(0.083)
Residual for Item e with Residual for Item h	0.306***	(0.085)
Panel C - PST Post-Survey - Coaching Satisfaction		
Sub-Factor: Support and Feedback (α = 0.969, IIC = 0.775)		
a. My clinical mentor helped me identify next steps to improve my teaching.	0.947***	(0.014)
d. My clinical mentor provided helpful coaching about presenting instructional content that helped me improve my teaching.	0.942***	(0.016)
e. My clinical mentor provided helpful coaching about planning instructional activities and materials that helped me improve my teaching.	0.914***	(0.023)
f. My clinical mentor provided helpful coaching about questioning students about instructional content that helped me improve my teaching.	0.839***	(0.035)
g. My clinical mentor explained how changing certain aspects of my teaching would improve student learning.	0.857***	(0.034)
c. When my clinical mentor observed and evaluated my teaching, I felt her/his evaluations were accurate.	0.820***	(0.045)
h. Overall, my clinical mentor's feedback helped me to improve.	0.904***	(0.025)
i. My clinical mentor observed me teach frequently enough.	0.880***	(0.030)
j. My clinical mentor provided me with feedback frequently enough.	0.923***	(0.020)
Sub-Factor: Autonomy and Encouragement (α = 0.952, IIC = 0.832)		
k. When I struggled with my teaching, I felt comfortable going to my clinical mentor for help.	0.940***	(0.017)
l. My clinical mentor's expectations of me were appropriate to my experience.	0.955***	(0.013)
m. My clinical mentor allowed me to make my own instructional decisions.	0.879***	(0.034)
n. I felt comfortable taking instructional risks in front of my clinical mentor.	0.864***	(0.037)
Covariance Structure		

Support & Feedback with Autonomy & Encouragement	0.922***	(0.026)
Residual for Item e with Residual for Item i	0.295**	(0.103)
Residual for Item f with Residual for Item g	0.552***	(0.084)
Residual for Item m with Residual for Item n	0.363**	(0.112)
Panel D - PST Post-Survey - Feeling of Preparedness - Teaching Skills		
Sub-Factor: Preparedness in Questioning Skills (α = 0.897, IIC = 0.634)		
b. Plan question sequences that help students develop deep conceptual understanding	0.813***	(0.042)
e. Ask questions that require students to discuss and/or write out their developing thoughts	0.779***	(0.041)
i. Develop questions that prompt students to grapple with the elements most necessary for understanding a text or concept	0.887***	(0.022)
j. Challenge students to wrestle with deep questions by providing adequate wait time	0.826***	(0.033)
k. Challenge all students by using strategies for calling on all students equitably	0.768***	(0.048)
Sub-Factor: Preparedness in Other Instructional Skills (α = 0.875, IIC = 0.539)		
a. Focus on essential information when presenting content		
c. Provide activities and materials that are relevant to students' lives	0.720***	(0.053)
d. Provide examples, illustrations, analogies, and labels for new concepts and ideas	0.771***	(0.048)
f. Plan activities that build curiosity	0.839***	(0.034)
g. Present content using visuals that establish the purpose of the lesson	0.788***	(0.035)
h. Incorporate multimedia, technology, and resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.)	0.857***	(0.032)
Covariance Structure	0.692***	(0.059)
Questioning with Other	0.911***	(0.026)
Panel E - CT Survey - Frequency of Coaching Practices		
Sub-Factor: Debriefing (α = 0.846, IIC = 0.524)		
b. Offer feedback on her/his teaching outside of class time when students were not present	0.621***	(0.050)
i. Share data or evidence about lessons you observed her/him teach	0.753***	(0.038)
j. Ask her/him reflective questions	0.761***	(0.037)
k. Analyze student work with her/him	0.677***	(0.046)
l. Share specific next steps for her/him to work on in order to improve her/his teaching	0.768***	(0.034)
Sub-Factor: Developing Practice (α = 0.697, IIC = 0.365)		
g. Model a specific instructional skill or move when students were not present	0.639***	(0.055)
h. Model a specific instructional skill or move for her/him during a lesson	0.449***	(0.074)
m. Use evaluation data to provide recommendations for improvement	0.730***	(0.041)

n. Provide opportunities outside of regular instruction to practice specific teaching moves	0.598***	(0.067)
Sub-Factor: Collaborative Coaching Practices (α = 0.660, IIC = 0.493)		
d. Co-plan a lesson or activity with her/him	0.706***	(0.066)
e. Co-teach a lesson or part of a lesson with her/him	0.636***	(0.064)
Covariance Structure		
Debriefing with Developing	0.955***	(0.033)
Debriefing with Collaborative	0.680***	(0.080)
Developing with Collaborative	0.772***	(0.073)
Residual for Item l with Residual for Item m	0.402***	(0.078)
Panel F: CT Survey - Coaching Frequency in Instruction Domain (α = 0.944, IIC = 0.607)		
a. Focus on essential information when presenting content	0.667***	(0.057)
b. Plan question sequences that help students develop deep conceptual understanding	0.762***	(0.038)
c. Provide activities and materials that are relevant to students' lives	0.833***	(0.030)
d. Provide examples, illustrations, analogies, and labels for new concepts and ideas	0.805***	(0.034)
e. Ask questions that require students to discuss and/or write out their developing thoughts	0.718***	(0.048)
f. Plan activities that build curiosity	0.814***	(0.037)
g. Present content using visuals that establish the purpose of the lesson	0.818***	(0.030)
h. Incorporate multimedia, technology, and resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.)	0.756***	(0.036)
i. Develop questions that prompt students to grapple with the elements most necessary for understanding a text or concept	0.826***	(0.039)
j. Challenge students to wrestle with deep questions by providing adequate wait time	0.777***	(0.050)
k. Challenge all students by using strategies for calling on all students equitably	0.847***	(0.027)
Covariance Structure		
Residual for Item a with Residual for Item i	0.437***	(0.074)
Residual for Item I with Residual for Item j	0.437***	(0.102)
Residual for Item j with Residual for Item k	0.426***	(0.079)