TN EDUCATION RESEARCH ALLIANCE

Improving Student Teachers' Feelings of Preparedness to Teach Through Recruitment of Instructionally Effective and Experienced Cooperating Teachers

A Randomized Experiment

Matthew Ronfeldt, Emanuele Bardelli, Hannah Mullman, Matthew Truwit, Kevin Schaaf, Julie C. Baker

Abstract

Prior work suggests that recent graduates from teacher education programs feel better prepared to teach and are more instructionally effective when they learned to teach with more instructionally effective cooperating teachers. However, we do not know if these relationships are causal. Even if they are, we do not know if it is possible to recruit cooperating teachers who are, on average, significantly more effective than those currently serving. This paper describes an innovative strategy to use historical administrative data on teachers to recommend the most instructionally effective and experienced teachers in various districts and subject areas to serve as cooperating teachers. In collaboration with a large teacher education program, partnering districts were randomized to receive either recommendation lists or use business-as-usual approaches. Those districts that received recommendations recruited significantly and meaningfully more effective and experienced cooperating teachers. Additionally, preservice student teachers who learned to teach in these same districts felt significantly better prepared to teach. This study offers an innovative and low-cost strategy for recruiting effective and experienced cooperating teachers and presents some of the first evidence that learning to teach with instructionally effective cooperating teachers has a causal impact on feelings of preparedness to teach.

WORKING PAPER
2019-06

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

Acknowledgements: We appreciate the generous financial support that was provided for this research by the Institute of Education Sciences (IES), U.S. Department of Education through the Statewide, Longitudinal Data Systems Grant (PR/Award R372A150015). Emanuele Bardelli and Hannah Mullman also received pre-doctoral support from the Institute of Education Sciences (IES), U.S. Department of Education (PR/Award R305B150012). We also appreciate comments on 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. 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.

Introduction

In order to receive initial certification, teacher candidates complete 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 CTs specifically – has important influences on preservice student teacher (PST) development. New research suggests that PSTs who were mentored by instructionally effective teachers are more instructionally effective themselves once employed. 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, school, district and program leaders may not know who the most instructionally effective teachers in local districts are due, for example, to data privacy laws and availability of evaluation data. Second, school and district leaders may be resistant to hand 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 a CTs' instructional effectiveness.

This study describes an initiative that aims to increase the overall instructional effectiveness of teachers serving as CTs. In particular, we test whether providing recommendations for which CTs to recruit, based upon existing administrative information, to district and teacher education program leaders both raises the average level of effectiveness of CTs and improves the quality of preparation for PSTs. Working with one large teacher education program (TEP) and the many partnering districts in which it places PSTs, we created an algorithm to identify the most instructionally effective and experienced teachers in the districts, subjects, and grade levels in which CTs were needed. We then randomly assigned 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 used 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, PSTs who learned to teach with this group of CTs also felt significantly better prepared to teach at the end of their clinical training (by 0.5-0.7 standard deviation units across measures).

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 learned to teach with more instructionally effective CTs during their preservice student teaching experiences. In Tennessee, Ronfeldt, Brockman, and Campbell (2018) linked evaluation data of PSTs to the evaluation data of their CTs and found that PSTs 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 also had better student achievement gains. In subsequent studies, Ronfeldt, Matsko, Greene Nolan, and Reininger (2018) found similar associations between PST and CT observation ratings in Chicago, while Goldhaber, Krieg, and Theobald (2018a) found similar associations between PST and CT achievement gains in Washington state. These studies provide evidence to support policies, like those in Tennessee, that set minimum teaching evaluation requirements for teachers to serve as CTs. Finding similar relationships across three different studies, labor markets, and sets of measures for instructional effectiveness also suggests that these associations may be picking up real 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 these 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 relationship between CT and PST instructional effectiveness is an understanding of *how* the former influences the latter. For this guidance, we turn instead to existing literature reviews on the research in teacher education (and specifically clinical education), consisting of primarily qualitative studies, typically self-studies, of individual programs. Based upon Gwenn (2008) and their reviews of the existing research, Grossman, Ronfeldt, and Cohen (2008) 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 to learn and practice for new teachers as well as provides feedback on their teaching efforts.[1]

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 emulate 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." As part of his Social Learning Theory, Albert Bandura (1976) has demonstrated the power of vicarious learning through observing others' model behavior. One might expect, then, that highly effective CTs are effective mentors because they model best practices that their PSTs are able to emulate and pick up in their own practice. Conversely, those CTs who model poor instruction might inadvertently pass less effective practices along to PSTs. Some literature suggests that PSTs have trouble discriminating when to appropriate the behaviors and teaching practices used by their CTs (Rozelle & Wilson, 2012), while the teacher socialization literature provides evidence that novices in clinical placements may actually be incentivized to pick up custodial, regressive teaching practices modeled by their CTs (Hoy & Woolfolk, 1990; Zeichner & Gore, 1990). Recruiting instructionally effective mentors – for example, with high observation ratings along measures of classroom culture -- could be an antidote to some of this socialization.

In terms of mentoring, rather than modeling, Schwille (2008) studied the strategies used by mentors of novice teachers who were known as effective embodiments of "educative mentoring" – mentoring grounded in learning theories that position the learner (here, the PST) as an active participant in the learning process. Given this dynamic, it is conceivable that more effective P-12 teachers are able to translate their considerable instructional skills into more effective mentoring for their PSTs. Schwille (2008) documented many such mentoring strategies, including providing 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. She contrasted these deliberate coaching practices with an "osmosis" approach "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 also provide both emotional support when needed (Glenn, 2006) and a balance of autonomy and encouragement (Yendol-Hoppey, 2007). It is possible that more instructionally effective teachers are more adept at

[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.

these mentoring practices; it is also likely that these mentoring practices require skills and capacities distinct from those needed for effective teaching of P-12 students.

The empirical basis linking CT practices – whether as mentors or as models – to outcomes for PST learning is especially thin. One exception is McQueen (2013) who designed a training program focused on supporting randomly assigned CTs in providing their candidates with more choice/autonomy about which area of teaching on which to focus and then maintaining a sustained focus in their feedback on that area over time. Per the typology described above, this training promoted a mentorship model for CTs, rather than a modeling one. McQueen found that those candidates who worked with trained CTs received stronger evaluations on their teaching, though differences were significant in only some specifications. These results are consistent also 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, Blazar, and Hogan (2018) for a review of this literature).

We know of two other studies that attempt to link CT practice with PST outcomes. Matsko, Ronfeldt, Greene Nolan, Klugman, Reininger, and Brockman (2018) looked at all CTs who served in the Chicago area and found evidence in support of the two functions of CTs – as models and mentors – that Grossman, Ronfeldt, and Cohen (2008) described. Namely, they found that PSTs reported feeling better prepared to teach at the end of their programs when CTs received better observation ratings (based upon the district evaluation rubric) and when PSTs reported that their CTs modeled more effective instruction as well as provided more frequent and/or better quality feedback, instructional support, autonomy and encouragement, collaborative coaching, and job assistance. In a subsequent, related study, Ronfeldt, Matsko, Greene Nolan, and Reininger (2018) also found that PSTs had better first-year observation ratings (based upon district evaluations) when their CTs received better observation ratings and when their CTs reported more mentoring focused on specific instructional practices, including those evaluated on the district rubric.

Based upon literature reviewed above, there is then evidence that CTs perform both a modeling and mentoring function and that both functions are related to PST outcomes. Less clear, though, is how these two functions could serve as mechanisms for explaining the relationship between CT and PST instructional effectiveness observed in prior studies. Finding measures of CT instructional effectiveness to predict PST instructional effectiveness and readiness to teach intuitively seems to support the first function – modeling effective instruction – as a mechanism for the observed relationship. Namely, PSTs may improve simply by observing and emulating the instructionally effective teaching being demonstrated by their CTs. However, it is also possible that coaching quality fully mediates any impact that CT instructional quality may have on PST instructional quality. More instructionally effective teachers of P-12 students may tend to be better instructional coaches of new teachers, perhaps because their knowledge of effective teaching supports them in being able to better notice and provide feedback on what may be lacking in their PSTs own enactments of practice. In other words, PST improvement may not result from emulating their CTs' teaching at all, but instead the improvement may result from rich opportunities to learn to teach that were deliberately designed and facilitated by more instructionally effective CTs who are better at designing these kinds of opportunities.

For mentoring to fully mediate the impact of CT instructional effectiveness on PST instructional effectiveness, though, we would first expect to find evidence that more instructionally effective teachers of P-12 students provide, on average, higher quality mentoring to PSTs. There is some evidence that this is the case. Ronfeldt, Goldhaber, Cowan and Bardelli (2018) developed and studied an initiative similar in many ways to the study described in this paper. After their partner programs over-recruited CTs, the authors used prior administrative data on CTs' instructional effectiveness (observation evaluations and student achievement gains) and years of experience, as

well as the schools in which they worked (average student achievement gains and teacher retention rates) to create a "placement index" meant to predict more and less promising placements. Using the median as the cutoff, the authors randomly assigned PSTs to be placed in either lower- or higher-index placements. Compared to PSTs assigned to low-index placements, peers placed in high-index placements reported that their CTs modeled or demonstrated more effective instructional practices; though not statistically significant, they also tended to feel better prepared to teach. More relevant to the present argument, high-index PSTs reported that their CTs engaged in more frequent coaching activities (including provision of feedback) that was also of a better quality; furthermore, they reported greater opportunities to practice different aspects of teaching in their placements.

While these results are consistent with the conclusion that more instructionally effective teachers of P-12 students tend, when serving as CTs, to provide better mentorship to PSTs, other explanations are possible. Beyond measures of instructional effectiveness, the placement index was also based on school-based measures, including average teacher retention, which is known to signal school working conditions (Ronfeldt, 2012). It is then possible that better school working conditions facilitated higher-index CTs in having more support and opportunity to mentor PSTs; that is, it could have been an effect of the placement school and not of the CT. The present study extends this prior work by focusing on only measures of teachers' instructional effectiveness and experience (without school-based measures) to predict more promising placements. It also randomizes neighboring districts to receive recruitment lists rather than randomizing PSTs to either higher- or lower-index placements. An advantage of the present randomization approach is that it allows us to have a business-as-usual comparison, whereas the comparison in the prior study was the low-index group which was, in some sense, manufactured as part of the research. This design allows us to test whether an intervention that is relatively low-cost and easy to reproduce can improve placements, on average, over typical approaches. In so doing, it also offers credibly causal evidence for the impact of being assigned to an instructionally effective and experienced CT on PST readiness to teach.

The current state of placement procedures

The review above suggests 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 student teaching placements are currently made, 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 placement procedures are already targeting and getting the most instructionally effective teachers to serve.

Existing placement procedures.

A handful of empirical studies help shed light on factors that influence the selection of CTs. In particular, demographic match between PST and CT, proximity to the TEP, and CT and placement school characteristics seem to influence which teachers get selected to serve (Krieg, Goldhaber, & Theobald, 2016; Maier & Youngs, 2009). We know of two studies that explore the placement procedures in specific labor markets. Reflecting the literature reviewed above, both of these studies find that TEP leaders and other stakeholders report considering a potential CT's ability to model effective instruction with students as well as support and mentor a PST. In the first of these studies, set in Washington state, St. John, Goldhaber, and Krieg (2018) identify a similar five-step process used across eight TEPs to make PST placements. The authors point out that day-to-day demands and concerns of different stakeholders may cause placement procedures to deviate from these steps, but that broadly, TEPs begin by assessing their needs and contacting district and

schools. The schools and districts evaluate their capacity to host, and eventually, PSTs, CTs, and principals meet to determine whether the placement is a good match.

Conversely, in Tennessee, where the present study takes place, Mullman and Ronfeldt (under review) found that placement 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 wanting full control of the process to allowing PSTs themselves to find their own placements. Moreover, if a TEP placed PSTs in multiple districts, they typically used a variety of systems for selecting CTs. Tennessee also has policies in place for clinical practice, including requirements for diversity of experience and a minimum of two placements, which gave TEPs additional considerations for selecting CTs.

In their National Council on Teacher Quality (NCTQ) report, Rickenbrode, Drake, Pomerance, and Walsh (2018) look across TEPs for graduate students and conclude that CTs' instructional effectiveness is not a consistent priority in placement procedures. Out of the 506 TEPs they study, they find that even in the eight states that set effectiveness criteria for CTs, only about half of programs take action to ensure these are honored and met. In the context of discussions of the current study, stakeholders have raised a variety of potentially competing priorities that might play a role in placement, including rewarding seniority, providing "help" to a struggling teacher, and a sense of 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 (under review) 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. While district leaders and school administrators might know who the most effective teachers are, they may not share that information with TEPs. Mullman and Ronfeldt (under review) 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 include differences in opinion about the role of the CT (i.e. as mentor or model), social networks (TEPs often recruit alumni from their programs to serve), ease of onboarding (once a TEP has a relationship with a teacher, they may try to use that CT again), or conceptions of how well a teacher works with other adults (St. John, Goldhaber, & Krieg, 2018; Mullman & Ronfeldt, under review). There is a prevalent belief, for example, 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. Moreover, PSTs who provide feedback about their satisfaction with their CT, whether solicited through an evaluation or not, might focus more on their perception of the quality of their relationship or the kinds of coaching they received, rather than on perceptions of value-added measures.

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, Krieg, & Theobald, 2018b; SAS Institute, 2014; Ronfeldt, Bardelli, Brockman, & Mullan, in press). In Tennessee, Ronfeldt and colleagues (in press) explored this possibility by measuring the impact of serving as a CT on both value-added measures and observation ratings. Allaying these concerns, they found no effects on value-added and small, positive effects on observation ratings.

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 hesitant to ask the same high-quality CTs to serve again and again as they worried they were putting undue burden on these teachers.

Do existing placement procedures work?

The wide variation in placement procedures – as well as the many obstacles to recruitment that exist – cast some doubt that the 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 valued-added to student achievement measures (VAMs) than other teachers in the state, though they had similar levels of teaching experience. Across preparation programs in Chicago, Gordon, Jiang, Matsko, Ronfeldt, Greene Nolan, and Reininger (2018) find that, compared with non-CTs, 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, Krieg, and Theobold (2018a) find that, all else being equal, teachers with more experience are more likely to host a student teacher, but teachers with greater VAMs are not.

Given that recruiters seem 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. Even if enough effective alternatives were available, we were concerned that some of the other obstacles described above, including concerns over evaluations being harmed and alternative recruitment criteria, 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 question 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 partnership 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 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; we decided to test whether providing better information alone could move the needle before moving to incentives.

Research Questions

RQ 1. Compared to the districts using business-as-usual recruitment strategies, are the CTs in districts that used the CT recommendation lists more instructionally effective?

RQ 2. Do student teachers report feeling more instructionally prepared when their CTs were recruited using targeted recruitment lists?

RQ 3. Does perceived mentoring quality change when CTs are recruited using targeted recruitment lists?

Methods

Research Design, Context, and Sample

For this initiative, we partnered with Tennessee Technological Institution (TTU), a large provider that uses a residency model, where candidates complete a year-long placement[2] in their CTs'[3] classroom(s). In 2017-18, the program placed 162 candidates in 22 neighboring districts. Candidates needed to complete their residency in subjects/grade levels appropriate for their specific program endorsement areas; e.g., candidates pursuing elementary endorsements were placed in grades K-5. Additionally, candidates were able to request a specific county/district in which they wanted to be placed, especially to accommodate geographic and travel constraints.[4]

We used this information to identify, for each candidate, all teachers that matched her requested county/district-by-grade band-by-subject "block." 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. 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 begin their recruitment with teachers named on the lists, 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 it inappropriate or unwise to recruit and to instead move to the next listed teachers. Among districts 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.

Recruitment Procedures

In this section, we elaborate on the specific algorithm we used to generate the recommendation lists. We calculate a composite measure of instructional effectiveness as the

[2] PSTs in TTU completed their residency experiences in only one placement, except for 20 PSTs had musical or special education placements, and needed to complete a second placement to fulfill their specific credentialing requirements. We generated new targeted recruitment lists for these PSTs and shared them with TTU following the same process as for the same 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/candidates. 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.

weighted average of observation ratings[5] (OR), value-added measures[6] (VAMs), and years of experience. We first standardize each measure within teaching field[7] at the state level. This procedure uses 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 placement block b. For OR and VAM, we average the scores for the three preceding school years, weighing the year immediately preceding placement as 50% of that measure and the other two 25% each. 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 calculate a final placement quality measure as

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

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 2016-2017.

Missing Evaluation Data

We have some missing evaluation data for the years that we use to calculate the placement quality measure. At this time, we decided not to impute or otherwise calculate possible values for these data (e.g., using Bayesian Shrinkage). Instead, we just remove the variable from the calculations and adjust the weights to reflect the data that is present. For example, if observation scores are not available for teacher i for time $t-3$, her evaluation scores will be weighed as 0.50 for $t-2$ and 0.50 for $t-1$. Other combinations of missing data follow 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.

---

[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 teaching fields as a proxy for endorsement area for teachers. We identify teaching 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 placements would fulfill the requirements for being recommended for a specific endorsement.

CT Eligibility

We decided that teachers are eligible to be CTs when their quality measures are in the upper three quintiles of the quality distribution. Thus, the targeted recruitment lists are organized by quality score, where the potential CTs with the highest quality score are at the top of the list and thereby the ones we ask the district/TEP leaders to recruit first. If the district/TEP leader asks all teachers on the lists we send them and still cannot recruit a CT for a candidate, at that point, we expect them to recruit in whatever way they typically would otherwise. Our rationale is that their business-as-usual approaches are preferable to suggesting they recruit CTs towards the bottom of the quality distribution. Additionally, we want to avoid the possibility of recommending CTs who may not meet the minimum level of effectiveness (LOE) score to serve as a CT.

This approach also mitigates a potential sensitivity that might arise with the practice of sending districts a ranked list of recommended teachers – because our list only includes teachers who are ranked approximately at or above "average" on our quality index, we could communicate to district partners two key messages: 1) Begin recruiting from the top of the list, because these are the most highly recommended teachers and (left unstated, but also important) there are in many cases large substantive differences between teachers at the top and bottom of the list, and 2) All the teachers on the list are recommended and there should be no stigma associated with being a teacher ranked near the bottom of the list.

Outcomes of Interest

The outcomes of interest for this paper are survey-based reports of feelings of preparedness, frequency of coaching, and satisfaction with coaching. We surveyed student teachers at the beginning (pre-survey) and at the end (post-survey) of their field placement. We also surveyed CTs once during the second half of field placement. Student teachers were surveyed about all outcomes while CTs were asked to report on frequency of mentoring practices. 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 fit indices for each model.

Feelings of Preparedness (PST Survey)

We measure feelings of preparedness in both pre- and post-surveys. We divided this construct into two correlated sub-constructs: readiness in questioning skills and readiness in other instructional skills. The first sub-factor includes five items that focused on readiness in developing, planning, and implementing questions to engage students in understanding a concept; we include a focus on this construct because the state has 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 focused on other aspects of planning and delivering instruction, such as developing materials, providing examples or analogies for new concepts, and using visuals during a lesson.

Mentoring Frequency (PST Survey)

We measure coaching frequency using four sub-constructs that focus on common mentoring practices, data-driven mentoring practices, collaborative coaching practices, and modeling coaching practices. Common mentoring practices include two items asking about the frequency of observations and of prompts to practice a specific aspect of teaching practice. We see these mentoring 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 measure 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, coaching about planning instructional activities, coaching about questioning students, explaining how certain changes to my practice would impact student learning) and the frequency of feedback (i.e., feeling that the mentor's evaluations were accurate, helpfulness of mentor's feedback, frequency of mentoring and feedback). The autonomy and encouragement sub-factor includes four items that measure the extent to which student teachers felt comfortable asking the CT for help or taking risks in front of the CT (i.e., feeling comfortable asking the mentor for help and to take risks in front of them, feeling that the mentor's expectations were appropriate, and feeling to have the ability to make independent instructional decisions).

Mentoring Frequency (CT Survey)

The CT survey included two main factors for mentoring practices: a general factor with three sub-factors and a specific factor on instructional practices. We divided the general factor on frequency of mentoring 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 student teacher 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 use the text from the domain descriptors from the TEAM rubric as question stems for this factor.

Analysis

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

$$Y_{ij} = \beta_0 + \beta_1 \cdot Treat_{ij} + \phi_j + \epsilon_{ij}$$

where Y_{ij} is the outcome of interest for CT or PST i in request field j, $Treat_{ij}$ is an indicator variable taking the value of 1 if CT i was recruited in a treated district. ϕ_j is a placement field fixed effect, and ϵ_{ij} are standard errors clustered at the district level. β_1 captures the treatment effect of receiving the targeted recruitment list on the outcome of interest.

We conduct three separate robustness checks to test the assumptions of our preferred model. 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 replace the placement field fixed effect with a field placement random effect. [8] Third, when using feelings of preparedness as an

[8] 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

outcome, we include pre-placement feeling of preparedness scores as a covariate in our models in order to control for possible imbalance in student teachers' initial feelings of preparedness (see Appendix Table 1). Overall, we find that the results of our preferred models are robust against these three checks.

Results

RQ 1. Placement Contrast

Table 1 summarizes the differences between CTs in districts receiving recruitment 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 (0.415 s.d. units), VAM scores (0.683 s.d. units), and years of experience (0.570 s.d. units).

We add indicators for placement field requests to increase the statistical power of these analyses and to account for possible differences between placement blocks, e.g., the possibility that secondary ELA teachers are rated higher (lower) on average than, say, elementary teachers. The estimates for these models are reported in the fourth row of Table 1. We find that the point estimates increase slightly, indicating that there are differences on evaluation scores between placement fields.

We use the average placement quality index to calculate the overall contrast between treatment and control CTs. This index allows us to compare CTs across teaching field as this variable is standardized within each placement field. We find that the placement quality for CTs in treatment districts is 0.425 standard deviations higher than the placement quality for CTs in control districts. This result is statistically significant at the 0.01 level. When we adjust these estimates for teaching field differences,[9] we find that the quality contrast increases to 0.476 standard deviation units.

RQ 2. Feelings of Preparedness

Table 2 reports the effect of being placed in a district that received the targeted recruitment lists on PSTs' feelings of preparedness. These results should be interpreted as an Intent-to-Treat (ITT) or reduced form estimate of the treatment effect of receiving a targeted recruitment list on student teachers' feelings of preparedness.

We find that PSTs in treatment districts report feeling significantly better prepared to teach by 0.621 standard deviation units (s.e. = 0.221, $p < 0.05$). We also see that these results are robust to how we calculate standard errors, to the inclusion of pre-placement controls, or to our decision to estimate field placement effects using fixed effects.

When we focus on the feelings of preparedness in specific subskills, we find that student teachers in treatment districts report feeling better prepared in both questioning skills (d = 0.658, s.e. = 0.223, $p < 0.05$) and other instructional skills (d = 0.583, s.e. = 0.221, $p < 0.05$), suggesting that the treatment effect is equally distributed across all teaching sub-skills that we measured.

---

that nested student teachers 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.

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

RQ 3. Reported Mentoring

Finding that candidates in treatment districts felt better prepared made us wonder: Why? One likely 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, teachers to serve as CTs; in turn, perhaps more effective and experienced teachers, on average, model better instructional practices thus helping the 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 candidates. To test this second possibility, we examined survey items related to the frequency and quality of coaching that candidates reported receiving and that CTs reported offering.

Results, which are summarized in Table 3, suggest that candidates 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 0.25 standard deviation units) in relation to data-driven mentoring practices. On the other hand, candidates in treatment districts tend 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 reported, differences between conditions are also mostly non-significant. That said, there were some notable trends: Treatment CTs report engaging in debriefing practices and coaching focused on the instructional domain more often and in developing practice activities less often than control CTs.

We use a mediation analysis to understand the relationship between reported mentoring practices and student teacher feelings of preparedness. We decompose our overall treatment effect in the direct path between the treatment indicator and the feelings of preparedness measure and an indirect path that links the treatment indicator to each mentoring practices measure individually and then a path between mentoring practices measures and feelings of preparedness (see Figure 1). We find that the inclusion of the measure of mentoring practices in the instruction domain explains about 11% of the overall treatment effect on feelings of preparedness. When we include measures of specific mentoring practices, we find that debriefing practices explain about 4% of the treatment effect, developing practices do not seem to impact the overall feelings of preparedness, and collaborative coaching practices seem to have a negative effect on the overall feelings of preparedness, decreasing the treatment effect by about 2%.

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 is to use 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. District leaders that used 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 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 doesn't seem groundbreaking that recruiters are able to recruit more instructionally effective and experienced teachers when told which are 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 already had prompting to target instructionally effective teachers as per state policy. Ronfeldt, Brockman, and Campbell (2018) showed that, even without recruitment 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. Could they do any better? Substantial doubts were also raised by stakeholders regarding evidence that instructionally effective teachers might be unwilling or unable to serve (Mullman & Ronfeldt, under review), at least in part because of local concerns that serving as CTs might harm evaluation scores (Goldhaber, Krieg, & Theobalrd, 2018b; Ronfeldt, Bardelli, Brockman, & Mullman, in press; 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 to serve as CTs in needed districts/subjects/grades. Given that recruiters were able to recruit 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.[10] Another possibility is that all recruiters knew who were the most instructionally effective and experienced teachers but instead used other criteria for their recruitment – e.g., 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, under review). More research is needed to understand why teachers at the top of the recruitment lists were not already being targeted.

The initiative also seemed to benefit PSTs, as candidates 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 candidates' sense of

[10] 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 mentors were recruited directly at the school level instead of relying on the central district office 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.

readiness 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 important to underscore that our focus on CTs' 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 of CTs and candidates' instructional effectiveness or feelings of preparedness to teach (Goldhaber, Krieg, & Thobald, 2018a; Matsko, Ronfeldt, Green, Nolan, Klugman, and Reininger, 2018; Ronfeldt, Matsko Greene Nolan, & Reininger, 2018; Ronfeldt, Brockman, & Cambpell, 2018). 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, Cowan, Bardelli, Johnson, and Tien (2018) find evidence that candidates 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 and positive effects on candidates' sense of readiness 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. Thus, the present study is the first, to our knowledge, to provide causal evidence that recruiting more effective and experienced CTs improves candidates' self-perceived readiness to teach. That said, in our view, helping candidates to feel better prepared is not enough, as it does not always predict becoming instructionally effective once in the classroom (Ronfeldt, Matsko Greene Nolan, & Reininger, 2018). 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.

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 or (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 through both mechanisms. When we examined whether or not PSTs in treatment districts reported better or more frequent coaching, results were mixed. Moreover, we found little evidence that better coaching practices explained or mediated the relationship we observed between treatment assignment and PST feelings of preparedness; the coaching practices we measured explained a small percentage of the effect of treatment on PSTs' feelings of preparedness but the main effects were still large and significant with the inclusion of the coaching measures. In other words, we found very little evidence in support for explanation (2) – that recruiting more instructionally effective CTs impacts PSTs' readiness to teach through improving the coaching that these PSTs are receiving 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 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. 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 observe but not attend to or learn from the aspects of instruction that they perhaps should be. Even if we did have adequate measures for PSTs' vicarious learning from CTs' instructional practice, we would need to do a similar mediation analysis as with Figure 1, but with PST vicarious learning measures, in order to determine the degree to which these may explain the main effects of treatment on PSTs' readiness to teach. More work is needed to identify the causal mechanism by which being assigned more instructionally effective teachers causes PSTs to feel more prepared to teach.

While it would be useful to know the mechanism (modeling, coaching, or some combination), 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' readiness 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 recruitment 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.

References

Bandura, A. (1976). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bollen, K. A. (1989). Structural equations with latent variables. New York: John Wiley Interscience.

Goldhaber, D., Krieg, J., & Theobald, R. (2018a). Effective like me? Does having a more productive mentor improve the productivity of mentees? CALDER working paper. Retrieved from 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. CALDER Working Paper.

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. Chicago, IL: University of Chicago Consortium on School Research.

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

Grossman, P., Ronfeldt, M., & Cohen, J. (2011). The power of setting: The role of field experience in learning to teach. In K. Harris, S. Graham, T. Urdan, A. Bus, S. Major, & H. L. Swanson (Eds.) *American Psychological Association (APA) Educational Psychology Handbook, Vol. 3: Applications to Teaching and Learning (pp. 311-334).*

Hoy, W. K., & Woolfolk, A. E. (1990). Socialization of student teachers. *American educational research journal*, *27*(2), 279-300.

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

Krieg, J., 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.

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.

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.

Matsko, K.K., Ronfeldt, M., Green Nolan, H., Klugman, J., Reininger, M., Brockman, S.L. (2018). Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education*. Advance online publication. DOI: 10.1177/0022487118791992.

Mullman, H., & Ronfeldt, M. (under review). The landscape of clinical preparation in Tennessee.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. Applied Psychological Measurement, 21(2), 173–184.

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?: Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis, 34(1), 3-26*.

Ronfeldt, M., Bardelli, E., Brockman, S., & Mullman, H. (in press). Will mentoring a student teacher harm my evaluation score? Effects of serving as a cooperating teacher on evaluation metrics. *American Educational Research Journal.*

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

Ronfeldt, M., Goldhaber, D., Cowan, J., Bardelli, E., Johnson, J., & Tien, C. D. (2018, April). Identifying promising clinical placements using administrative data: Preliminary results from ISTI placement initiative pilot. CALDER Working Paper. Retrieved from https://caldercenter.org/sites/default/files/WP%20189.pdf

Ronfeldt, M., Matsko, K.K., Greene Nolan, H., & Reininger, M. (2018). Who knows if our teachers are prepared? Three different perspectives on graduates' instructional readiness and the features of preservice preparation that predict them (CEPA Working Paper No.18-01). Retrieved from Stanford Center for Educational Policy Analysis: http://cepa.stanford.edu/wp18-01.

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.

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

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), Latent variables analysis: Applications for developmental research. Thousand Oaks, CA: Sage Publications, Inc.

Schwille, S. A. (2008). The professional practice of mentoring. *American journal of education*, *115*(1), 139-167.

St. John, E., Goldhaber, D., & Krieg, J., Theobald, R. (2018). How the match gets made: Exploring candidate placements across teacher education programs, districts, and schools. CALDER Working Paper. Retrieved from https://caldercenter.org/sites/default/files/CALDER%20WP%20111018.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). *New York, NY: McMillan.*

Tables

Table 1. Contrast of CT Quality Measures between Treated and Control Placements

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

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

Table 2. Student Teacher Post-Survey Differences between Treatment and Control Districts

	(1) Preferred Model	(2) Bootstrapped S.E.	(3) Pre-Survey Control	(4) R.E. Model
Feeling of Prep - Teaching Skills	0.621*	0.621*	0.492*	0.579***
	(0.221)	(0.265)	(0.198)	(0.144)
Readiness in Questioning Skills	0.658**	0.658**	0.526*	0.616***
	(0.223)	(0.246)	(0.193)	(0.154)
Readiness in Other Instructional Skills	0.583*	0.583**	0.457*	0.542***
	(0.221)	(0.211)	(0.208)	(0.136)

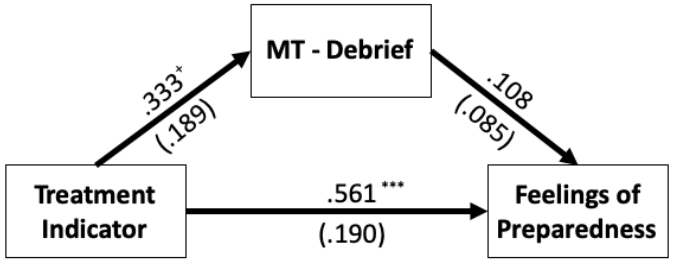
Notes. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Student Teacher Post-Survey Differences between Treatment and Control Districts

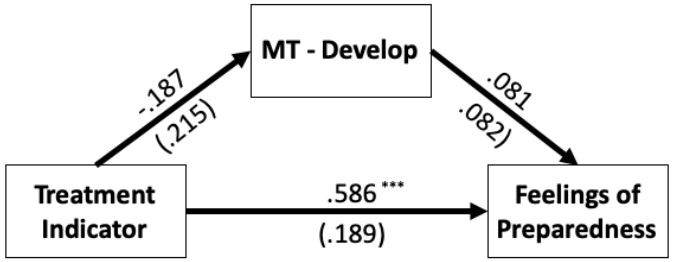
	(1) Preferred Model	(2) Bootstrapped S.E.	(3) Pre-Survey Control	(4) R.E. Model
Student Teacher Surveys				
Mentoring Frequency	0.181	0.181	0.182	0.169
	(0.147)	(0.170)	(0.201)	(0.110)
Common Mentoring Practices	0.143	0.143	0.151	0.145
	(0.184)	(0.209)	(0.231)	(0.100)
Data-Driven Mentoring Practices	0.236	0.236	0.282	0.242*
	(0.201)	(0.227)	(0.267)	(0.106)
Collaborative Coaching Practices	0.205+	0.205	0.152	0.166
	(0.111)	(0.149)	(0.153)	(0.163)
Modeling Coaching Practices	0.141	0.141	0.144	0.151
	(0.186)	(0.217)	(0.260)	(0.118)
Coaching Satisfaction	-0.142	-0.142	-0.156	-0.202+
	(0.170)	(0.214)	(0.266)	(0.109)
Support and Feedback	-0.180	-0.180	-0.176	-0.235*
	(0.169)	(0.201)	(0.265)	(0.112)
Autonomy and Encouragement	-0.104	-0.104	-0.136	-0.169
	(0.177)	(0.211)	(0.274)	(0.109)
CT Surveys	0.181	0.181	0.182	0.169
Frequency of Mentoring Practices				
Debriefing	0.326	0.326	0.249	0.348*
	(0.248)	(0.284)	(0.339)	(0.138)
Developing Practice	-0.150	-0.150	0.080	-0.235
	(0.175)	(0.245)	(0.315)	(0.225)
Collaborative Coaching Practices	-0.005	-0.005	0.045	0.038
	(0.238)	(0.238)	(0.350)	(0.114)
Coaching Frequency in Instruction Domain	0.180	0.180	0.372	0.105
	(0.170)	(0.201)	(0.245)	(0.213)

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

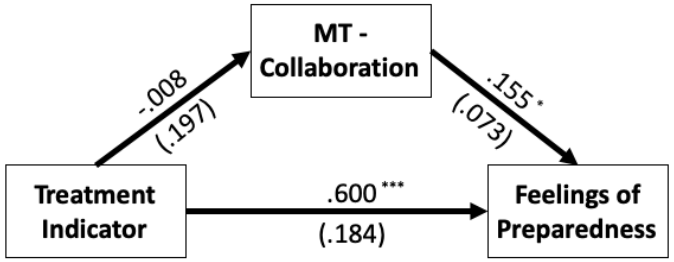
Figure 1. Mediation Models between Feelings of Preparedness and Mentoring Practices



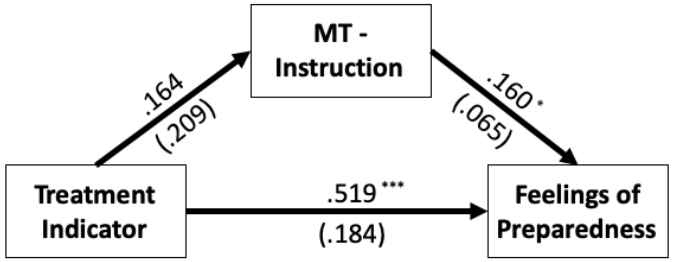
Total effect: .586 (.189), p= .002
Reduction: 4.1%



Total effect: .586 (.189), p= .002
Increase: 0.1%



Total effect: .586 (.189), p= .002
Reduction: 2.5%



Total effect: .586 (.189), p= .002
Reduction: 11.3%

Appendix Table 1. Student Teacher Pre-Survey Differences between Treatment and Control Districts

	(1) Preferred Model	(2) Bootstrapped S.E.	(3) R.E. Model
Feeling of Preparedness - Teaching Skills	0.317	0.317	0.185
	(0.255)	(0.289)	(0.130)
Readiness in Questioning Skills	0.319	0.319	0.236*
	(0.263)	(0.273)	(0.118)
Readiness in Other Instructional Skills	0.314	0.314	0.151
	(0.252)	(0.244)	(0.140)

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

Appendix Table 2. Correlation among PSTs' Coaching Satisfaction and CTs' Coaching Frequency Measures

	Frequency of Mentoring Practices			**Coaching Frequency in Instruction Domain**
	Debriefing	Developing Practice	Collaborative Coaching Practices	
Support and Feedback	0.217	0.115	0.279*	0.147
	(0.081)	(0.356)	(0.024)	(0.238)
Autonomy and Encouragement	0.223	0.117	0.300*	0.153
	(0.073)	(0.349)	(0.015)	(0.221)

Notes. This table reports pairwise correlations between measures of PSTs' satisfaction with coaching and CTs' coaching frequency. Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix 1 – 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.

We now cover each latent factor that we calculated. We first report the observed reliability estimates (i.e., alpha score and inter-item correlation) for each latent sub-construct within each measure. Then, we report the results of a CFA confirming the factor structure for each item. This includes fit indices, factor loadings, correlations between latent-subfactors, and error covariance terms. Finally, we report latent reliability estimate in the form of the Raykov's rho coefficient (Raykov, 1997).

Technical Appendix Table 1 – Fit Indices for Each Measurement Model

Fit Indexx		Pre- Prep Subskills	Post-MT Frequency	Post-MT Quality	Post-Prep Subskills	MT - Frequency	MT - Freq subskills
Chi Squared	Value	48.944	62.492	77.798	51.061	55.649	62.450
	df	42	45	61	43	40	41
	p	0.214	0.043	0.072	0.186	0.051	0.017
RMSEA	Lower Bound	0.026	0.052	0.044	0.037	0.045	0.064
	Estimate	0.028	0.052	0.082	0.053	0.048	0.114
	Upper Bound	0.075	0.103	0.123	0.108	0.095	0.165
	CFI	0.993	0.983	0.988	0.989	0.975	0.966
	TLI	0.991	0.976	0.984	0.985	0.966	0.954
	SRMR	0.038	0.056	0.061	0.038	0.053	0.078
	CD	0.962	1.000	0.994	0.975	0.895	0.944