JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2023.2287615

Routledge
Taylor & Francis Group

INTERVENTION, EVALUATION, AND POLICY STUDIES

Mentors Matter Recruitment Replication & Extension: Investigating Effects Across Implementation Years

Matthew Ronfeldt[a], Emanuele Bardelli[b], Matthew Truwit[a], Kevin Schaaf[c] and Julie Baker[d]

[a]Educational Studies, University of Michigan, Ann Arbor, Michigan, USA; [b]Annenberg Institute at Brown, Providence, Rhode Island, USA; [c]Tennessee Department of Education, Nashville, Tennessee, USA; [d]College of Education, Tennessee Technical University, Nashville, Tennessee, USA

ABSTRACT

The Mentors Matter Recruitment initiative leveraged state administrative data to recommend and successfully recruit more instructionally effective and experienced teachers to serve as clinical mentors, which, in turn, increased pre-service teachers' readiness to teach. Prior results, however, focus on self-reported outcomes and stem from a single field experiment, raising questions about their replicability. In this paper, we replicate the previous study with a second cohort, finding again that the initiative led to the recruitment of more instructionally effective and experienced mentors. In addition, we examine new outcomes from administrative and program data, observing that candidates assigned to mentors recruited through our lists were rated as significantly more instructionally effective during student teaching. Given these promising results, we tested the same initiative at three new programs under less oversight from research partners and policymakers, finding that even under these more typical conditions, the initiative successfully produced significantly more instructionally effective and experienced mentors for a third time, though with smaller effects.

ARTICLE HISTORY
Received 28 July 2022
Revised 5 October 2023
Accepted 31 October 2023

KEYWORDS
Clinical mentors; experiment; replication; teacher preparation

Since the introduction of teacher education, field (clinical) experiences have been a cornerstone of teacher preparation. Field experiences provide pre-service student teachers (PSTs) opportunities to experiment with enacting teaching practice in a classroom under the supervision of an accomplished teacher, whom we refer to as a clinical mentor (CM), also commonly known as a cooperating or mentor teacher. CMs have long been assumed to play critical roles in cultivating the pedagogical skills of PSTs, serving both as exemplars of effective teaching practice and as coaches who help PSTs improve through observation and feedback on their teaching during field experiences (including student teaching). However, only recently have researchers started to empirically test which CM characteristics are associated with improved PST outcomes (see, Ronfeldt, 2021, for a review of this literature).

CONTACT Matthew Ronfeldt ronfeldt@umich.edu Educational Studies, University of Michigan, Ann Arbor, Michigan, USA.

Ten studies over the past decade have found large-scale correlational evidence of a relationship between the instructional effectiveness of CMs and that of the PSTs that they mentored (Bastian et al., 2020; Goldhaber et al., 2020, 2021, 2022; Matsko et al., 2022; Ronfeldt et al., 2013, 2020, 2021; Ronfeldt et al., 2018a; Ronfeldt et al., 2018b). Spanning four states, a variety of regression modeling approaches, and numerous measures for instructional effectiveness—including observation ratings, value-added to student achievement measures (VAMs), and self-reported (survey-based) measures—these studies have all found significant and positive associations.

Though these consistent correlational findings could suggest that instructionally effective CMs *cause* their PSTs to become more instructionally effective, other explanations are possible. For example, more promising and motivated PSTs, who likely would have become more instructionally effective regardless of who served as their CMs, may seek out or select more instructionally effective CMs. This selection of PSTs to CMs (or vice versa) could present non-causal, endogenous, and alternative explanations for the observed relationships among these prior studies and has motivated the pursuit of credibly causal evidence from even more recent randomized control trials (RCTs).

In the first of these RCTs, the Improving Student Teaching Initiative (ISTI), Ronfeldt et al. (2018b) asked two partner teacher preparation programs to recruit twice as many potential CMs as needed while continuing to meet previous standards for high-quality placements. The authors then created two lists of potential field placements for PSTs in each program—one they predicted to contain the more promising group of placements and the other the less promising group—based upon historical administrative data for the characteristics of CMs (VAM scores, observation ratings, years of experience) and their schools (average teacher turnover rates, school VAMs) shown in prior research to predict better employment rates and observation ratings for PSTs. PSTs were then randomly assigned to one list or the other.

The authors found that PSTs assigned to the placements they predicted to be more promising both reported more opportunities to learn to teach and felt more ready to teach at the end of their field placements. Moreover, these same PSTs also reported feeling that their CMs were more instructionally effective teachers and that their school placement sites had better working conditions. In a follow-up study, Goldhaber et al. (2022) analyzed PSTs' instructional effectiveness over the course of their field placement experiences, as measured by clinical assessment scores (observational ratings of instruction during student teaching). They found that PSTs assigned to the promising placements also demonstrated faster growth.

ISTI presented some of the first evidence that more instructionally effective and experienced CMs caused an increase in PSTs' instructional effectiveness. However, the experiment's placement lists drew on field placement school information in addition to CM characteristics, making it possible that differences between placement schools, rather than CMs, could explain the observed results.

The second of these RCTs—the Mentors Matter Recruitment (MMR) experiment—addressed these limitations by focusing solely on CM characteristics in the assignment of field placements. In this study, school districts working with one educator preparation program (EPP) were randomly assigned to receive "recommendation lists" which suggested the most instructionally effective and experienced teachers available to be

invited to serve as CMs. These recommendation lists ranked potential CMs in any given school district and grade/subject area using teacher evaluation scores (observation ratings and VAMs, referred to as TVAAS in Tennessee) and years of teaching experience. Ronfeldt et al. (2020) reported that "treatment" school districts—those that received the recommendation lists—were able to recruit CMs with significantly greater observation ratings, TVAAS, and experience as compared to school districts that followed business-as-usual CM selection strategies. Consequently, PSTs in the treatment condition reported feeling significantly more ready to teach at the end of their clinical placements than those in the control condition.

The emerging experimental evidence seems to suggest that increasing the instructional quality of CMs directly results in improved field placement experiences and more capable and ready PSTs. However, both RCTs were conducted in partnership with an EPP that had a history of successful prior collaboration with both the TDOE and the lead author and therefore likely represented something of an ideal partner in terms of its commitment to partnership, research, and improvement efforts. This raises concerns about the replicability of results to other, more typical, EPPs. In addition, to this point, MMR focused only on PSTs' self-reported readiness to teach rather than their observed instructional effectiveness or other workforce outcomes. As prior evidence suggests that feeling better prepared is uncorrelated with actually being more instructionally effective (Ronfeldt et al., 2021), we felt it important to go beyond the self-reported measures used in the original MMR study.

There is also rising emphasis among scholars regarding the need for replication of experimental results in education, especially given many recent failed efforts at reproduction. Replication under both the same and different experimental designs allows for a clearer delineation between spurious one-off effects and concrete causal relationships that scale up beyond the original study (Makel & Plucker, 2014; Steiner et al., 2019; Simpson, 2022; Wong & Steiner, 2018). Re-implementation and replication across a variety of settings and contexts can also help in determining which experimental results may be contextual anomalies and which might generalize beyond the original population.

In this paper, we describe a pair of studies that build upon the "original" MMR evaluation (Ronfeldt et al., 2020) in two different ways. In the "replication study," we re-implement the original experiment (Cohort 1, 2018–19) with a different cohort of PSTs and CMs within the same EPP (*Cohort 2*, 2019–20). We use the same experimental design, implementation, and outcomes but also consider three new outcomes—clinical assessment scores (observational measures of PSTs' instructional effectiveness during student teaching), employment after program completion, and first-year observation ratings. While we were able to fully replicate the initial implementation of our original experiment involving the development of recommendation lists and the recruitment of mentors, we note that the COVID-19 pandemic, which began impacting schools in March 2020, substantially impacted both preservice preparation and later elements of data collection for *Cohort 2*, which we further detail in the *Measures* section below.

In the "extension study," we partnered with three new programs (2020–21) that varied in size and geographical location and asked them to recruit all CMs using our recommendation lists. Leveraging four years of prior CM data from these programs, we estimate an interrupted time series to assess whether having access to the

recommendation lists resulted in their recruitment of more instructionally effective and experienced CMs as compared with prior years. Here, we note again that the pandemic coincided with the implementation of our study, where only in the year of our intervention were teachers being recruited and choosing to serve as mentors during a pandemic, the implications of which we unpack below.

The replication study is a direct experimental replication with the goal of assessing estimate stability over similar re-implementation conditions and includes new outcome measures; the extension study is a quasi-experimental scale-up effort with a set of new EPP partners who adopt the initiative in their own ways, providing estimates of real-world implementation effects of the MMR initiative beyond the ideal case in the first experiment. Both studies serve to ensure that the results of the original MMR experiment are valid, meaningful, and generalizable before encouraging their use to inform policy implementation.

The same questions that guided Ronfeldt et al. (2020) also guide this paper: (RQ1) Do CMs in districts randomized to receive recommendation lists have higher average effectiveness scores and experience compared with those in districts following business-as-usual recruitment strategies? (RQ2) Do PSTs report feeling more instructionally ready when their CMs were recruited using recommendation lists? In addition, this paper asks two new questions: (RQ 3) Do PSTs have better clinical evaluations (of their student teaching performance), employment rates, or first-year observation ratings when their CMs were recruited using recommendation lists? and (RQ 4) When we scale up this intervention with a new set of EPPs, do we observe a similar contrast in mentor instructional effectiveness and experience?

In the replication study (RQs 1–3), we find largely consistent experimental effects to the original RCT, where having access to the recommendation lists led to the recruitment of substantially more instructionally effective and experienced CMs. Across both years of implementation, PSTs assigned to treatment mentors reported feeling significantly more prepared to teach. Notably, *Cohort 2* effects were about half the magnitude as those for *Cohort 1* and not statistically significant; however, as we administered the survey for *Cohort 2* in the spring of 2020 during the first wave of the COVID pandemic, PSTs were faced with much uncertainty and often had to move to online instruction, which may explain some of these between-cohort differences. In terms of new outcomes, we find that PSTs in treated districts received significantly higher clinical assessment scores. They had statistically similar employment rates and first-year observation ratings, though point estimates trended positive.

In the extension study (RQ 4), we find that receiving recommendation lists increased the average instructional quality and experience of recruited CMs, suggesting that the initiative has promise across various types of program contexts and with less support from research and policy partners. However, the magnitude of the improvement was smaller overall and appeared to depend on the average level of mentor instructional quality obtained via programs' business-as-usual CM recruitment strategies during the pretreatment period. Together, these two replication studies suggest that providing recommendation lists to teacher education programs is a viable, stable, and scalable way to improve the average instructional quality of the CM cohort and perhaps even the instructional readiness of their PSTs.

Methods

The Mentors Matter Recruitment (MMR) initiative was a set of three studies beginning in the 2018–2019 academic year. MMR started with the initial field experiment, summarized above (see Ronfeldt et al., 2020, for more detail) and was followed in consecutive years by the replication and extension studies described next. The first two field experiments (the "original" and "replication" studies) were implemented in collaboration with the Tennessee Department of Education (TDoE) and Tennessee Tech University (TTU). A third, quasi-experimental ("extension") study was implemented in collaboration with three new EPPs: Milligan University, Trevecca University, and the University of Tennessee at Martin.

The Original and Replication Field Experiments

Design

These RCTs sought to randomly assign the school districts that partner with TTU to either use business-as-usual recruitment procedures (control) or have their recruiters receive recommendation lists identifying the most instructionally effective and experienced teachers to target during CM recruitment (treatment). Lists were designed by ranking the most promising potential CMs in requested placement blocks—a grade level and/or subject request in a specific county—using teachers' prior observation ratings, TVAAS scores, and years of experience. We used up to three prior years of administrative data with teacher evaluation and experience information to create a composite measure that we call the recommendation index. Administrative data from the year prior to recruitment contributed up to 50% of the recommendation index scores, and data from two and three years prior each contributed 25%. Within each year, observation ratings and TVAAS each contributed 40% to the recommendation index, while teacher experience contributed the final 20%.[1]

State partners advised recruiters in treatment districts to use the lists by first trying to recruit the teacher at the top (i.e., the most instructionally effective and experienced) before moving to the second person next, etc. However, recruiters were assured that these were recommended and not required lists; if they had a good reason to skip someone (e.g., a teacher had too many other responsibilities; a teacher got poor reviews as a mentor previously), then they should move to the next person on the list.

Figure 1 summarizes the logic model for the original and replication field experiments. We hypothesized that recruiters in districts randomly assigned to receive these lists would recruit more instructionally effective and experienced teachers to serve as CMs. As a result, we hypothesized that PSTs assigned to these CMs would—after observing higher quality teaching modeled by their CMs and receiving higher quality coaching—be rated as more instructionally effective on their clinical assessments (i.e., observation ratings during student teaching). In addition, we hypothesized that they

[1]Since we developed the recruitment index based upon Ronfeldt et al. (2020), see this prior study for more details, including formulas used to calculate it and how missingness is handled. Of note, TVAAS are only available for teachers who teach in tested grade levels and subjects (in general, 3rd through 8th grade and selected high school courses). The calculations for the recommendation index for teachers who do not have a TVAAS score only include observation ratings and years of experience, reweighing the index 67% OR and 33% years of experience.

Condition	Description of Condition	Expected Outcomes: Beginning of Clinical Placements	Expected Outcomes: During Clinical Placements	Expected Outcomes: End of Clinical Placements	Expected Outcomes: Subsequent Year Employment	Expected Outcomes: Subsequent Year (*Cohort 1* Only)
Experimental	Districts randomized to receive recommendation lists for recruiters to use to recruit clinical mentors (CMs)	Experimental districts recruit more instructionally effective and experienced teachers to serve as CMs	Preservice student teachers (PSTs) receive stronger observation ratings during student teaching	PSTs report feeling more prepared to teach & more frequent/better quality mentoring from their CMs	PSTs are more likely to find employment (in part because they are more prepared to teach)	PSTs have better first-year observation ratings
Control	Districts randomized to use business-as-usual approaches to recruit clinical mentors (CMs)	Control districts recruit less instructionally effective and experienced teachers to serve as CMs	PSTs receive weaker observation ratings during student teaching	PSTs report feeling less prepared to teach & less frequent/worse quality mentoring from their CMs	PSTs are less likely to find employment (in part because they are less prepared to teach)	PSTs have better first-year observation ratings

Figure 1. Logic model.

would report on end-of-program surveys that they felt more prepared to teach and that they received more frequent and higher quality mentoring. Finally, we hypothesized that these same PSTs would be more likely to gain employment and be more instructionally effective (as measured by first-year observation ratings) in the subsequent year.

Sample

In the original study, 12 districts were randomly assigned to receive recommendation lists (i.e., treatment) while 10 were asked to use business-as-usual procedures (i.e., control); in the replication study, 15 districts were randomly assigned to treatment and 17 to control. Given some districts were assigned to the same condition (e.g., treatment) both years while others were not (e.g., control for *Cohort 1* and treatment for *Cohort 2*), we consider effect estimates for these various permutations below.

Across the two years of implementation, 315 PSTs participated—155 in *Cohort 1* and 160 in *Cohort 2*. Table 1 reports sample statistics for selected PST and CM characteristics. Most PSTs are White women; they have, on average, a GPA of 3.56, an ACT score of 22.8, and a Praxis core score of 168.7. Overall, CMs who participated in the initiative, irrespective of treatment condition, are effective teachers; they have, on average, an observation rating of 4.32, a TVAAS score of 0.58, and about 15.5 years of experience. This makes CMs about half a standard deviation more instructionally effective than the average teacher across the state, based on both ORs and TVAAS; they also are about 0.4 standard deviations more experienced. Finally, both PST characteristics and CM characteristics remained qualitatively similar across cohorts, with only minor variation.

Balance Check

Given randomization occurred at the district level, we checked for balance in our replication study sample on district-average K-12 student characteristics and teacher (i.e., potential CM) evaluation information in Table 2. We include teacher characteristics both at the district mean and the 90th percentile, given that CMs are typically recruited from the top of the distribution of teacher effectiveness. Across all student and teacher characteristics, we found no significant differences between treatment and control

Table 1. Pre-teacher and clinical mentor descriptive statistics.

	Combined			Cohort 1			Cohort 2		
	Mean	Std. Dev.	*N*	Mean	Std. Dev.	*N*	Mean	Std. Dev.	*N*
Panel A. PST characteristics									
Woman	0.868	0.325	290	0.905	0.257	130	0.838	0.370	160
White	0.975	0.147	290	0.990	0.058	130	0.963	0.191	160
Current GPA	3.560	0.347	290	3.503	0.228	130	3.606	0.415	160
ACT	22.81	2.626	282	22.57	2.071	128	23.00	3.003	154
Praxis	168.7	9.556	242	169.6	6.420	128	167.7	12.10	114
Panel B. CM Characteristics									
Recommendation quintile	3.913	1.321	300	3.948	1.328	155	3.876	1.317	145
Recommendation list rank	26.64	44.19	277	22.35	34.10	130	30.42	51.31	147
Recommendation index	0.504	0.630	300	0.527	0.634	155	0.479	0.627	145
Observation ratings (Std)	0.496	0.676	300	0.495	0.718	155	0.499	0.630	145
TVAAS	0.576	0.948	152	0.625	0.878	73	0.531	1.012	79
Years of experience (Std)	0.442	1.035	300	0.476	1.088	155	0.405	0.978	145
Observation ratings	4.318	0.380	300	4.294	0.404	155	4.344	0.352	145
Environment domain	4.661	0.335	282	4.614	0.345	143	4.709	0.319	139
Instruction domain	4.177	0.405	282	4.142	0.416	143	4.213	0.392	139
Planning domain	4.282	0.447	280	4.234	0.451	141	4.331	0.439	139
Professionalism domain	4.505	0.432	282	4.496	0.476	143	4.515	0.383	139
Level of effectiveness	4.542	0.611	260	4.676	0.522	117	4.434	0.656	143
Years of experience	15.51	9.377	300	15.90	9.872	155	15.08	8.833	145

Table 2. Balance check on placement district characteristics.

	(1) Mean	(2) Control mean	(3) Treatment mean	(4) Diff	(5) Effect size	(6) *p*-Value
Panel A. Student characteristics						
% African American/Black	12.915	5.169	5.391	0.222	0.03	0.947
% Hispanic/Latino	6.751	6.031	6.373	0.342	0.061	0.893
% Asian	1.304	0.953	0.713	−0.240	0.303	0.504
% Native American	0.371	0.328	0.364	0.036	0.247	0.586
% White	78.524	87.398	87.047	−0.351	0.029	0.949
% Hawaiian/Pacific Islander	0.135	0.128	0.110	−0.019	0.196	0.665
% Free or reduced price lunch	37.071	37.469	39.518	2.049	0.207	0.648
% Students with disabilities	14.952	14.578	14.681	0.104	0.057	0.900
% English language learners	0.480	0.334	0.387	0.053	0.145	0.748
Panel B. Teacher characteristics						
Observation ratings (std)	−0.029	−0.145	−0.098	0.047	0.122	0.764
Instruction domain (std)	−0.097	−0.179	−0.124	0.055	0.162	0.698
Environment domain (std)	−0.090	−0.145	−0.086	0.060	0.17	0.682
Planning domain (std)	−0.071	−0.131	−0.019	0.112	0.275	0.511
Professionalism domain (std)	−0.040	−0.126	0.020	0.147	0.383	0.362
TVAAS	0.034	−0.043	−0.158	−0.116	0.435	0.290
Years of experience	11.521	11.845	11.959	0.115	0.056	0.890
Observation ratings (90th pct)	0.868	0.932	0.989	0.057	0.147	0.718
Instruction domain (90th pct)	0.853	0.861	1.070	0.208	0.533	0.209
Environment domain (90th pct)	0.733	0.838	0.851	0.014	0.056	0.892
Planning domain (90th pct)	0.803	0.825	0.977	0.151	0.376	0.371
Professionalism domain (90th pct)	0.847	1.012	0.947	−0.065	0.279	0.505
TVAAS (90th pct)	0.913	0.924	0.756	−0.168	0.406	0.323

Notes. This table compares student and teacher characteristics of school districts that were offered and were not offered recommendation lists (i.e., treatment and control groups respectively). Effect sizes are calculated as the covariate-adjusted mean difference divided by the unadjusted pooled within-group SD. Joint chi square test for Panel A: $\chi^2(9) = 3.61$, $p = .93$; Panel B: χ^2 (13) = 18.03, $p = .16$.

districts, suggesting that our randomization was again successful in the second year of implementation (for student characteristics, $X^2(9) = 3.61$, $p = .93$; for teacher characteristics: $X^2(13) = 18.03$, $p = .16$; see Ronfeldt et al. (2020) for *Cohort 1* balance checks, which also suggest randomization was successful).

Measures

We focus on five sets of focal outcomes. The first two were included in the original study: PSTs' and CMs' self-reported survey measures. The next three move beyond self-report and are new to the replication study: PSTs' clinical assessment scores, employment rates, and observation ratings. We point out that the availability of each of these outcomes, as well as what they measure, were impacted by the pandemic in many ways, particularly for *Cohort 2*. We have outcomes from both cohorts across all measures except PSTs' observation ratings which, due to the COVID pandemic halting the evaluation of in-service teachers, are available only for *Cohort 1*. Even though we have data for the other measures, the following measures were collected during pandemic-affected months and so may have impacted the sample and/or the scores for one or both of the cohorts: PST post-survey (*Cohort 2*), CM survey (*end of Cohort 2*), PST clinical assessments (*end of Cohort 2*), and PST employment (*Cohort 2* and possibly *Cohort 1* members who were still seeking employment a year after graduation). We describe each these measures in further detail below.

Pre-Service Teacher Survey-Based Measures

We administered surveys to measure PSTs' impressions of their year-long clinical experiences (Ronfeldt et al., 2020 for more information about the measures and their psychometric properties). All PSTs participating in the MMR initiative were invited through email to fill out surveys before and after their clinical placements; we respectively call these the pre- and post-survey administrations. Administrative staff from TTU also contacted them to remind them to complete the survey before each data collection window closed.

We conducted confirmatory factor analyses to develop aggregate measures for each of the three survey-based outcomes of interest. All factors displayed good to excellent psychometric properties, justifying our factor structure.[2] *Mentoring frequency* measures the relative frequency of various mentoring practices during an average week of their clinical placements as reported by PSTs. These mentoring activities fall along four categories: (1) common mentoring practices, (2) data-driven mentoring practices, (3) collaborative coaching practices, and (4) modeling coaching practices. *Coaching satisfaction* measures PSTs' impressions of the extent to which they felt supported and coached by their CMs and of the level of autonomy and encouragement they were given during their clinical placements. *Feelings of readiness* measures PSTs' perceived readiness in specific teaching skills; we developed two sub-measures to this factor for questioning skills and other instructional skills.

Clinical Mentor Coaching Survey Measures

We surveyed CMs in nine different survey groups staggered throughout PSTs' clinical experiences to capture the full spectrum of mentoring practices that might vary over time. Each mentor was randomized within treatment condition into a survey group and asked to complete the survey considering their mentoring practices during the prior week.

CM surveys included items about the frequency of various general mentoring practices and the specific frequency of coaching in instructional domains aligned with the state's teacher evaluation rubric. As with the PST survey, we conducted confirmatory

[2]The details of these analyses are reported as a Technical Appendix to Ronfeldt et al. (2020).

factor analyses to develop aggregate measures for these survey-based outcomes. More precisely, we developed a general *mentoring frequency* factor with three subfactors and a specific factor on coaching around instructional practices. The general factor contained three correlated subfactors: debriefing, developing practice, and collaborative coaching practices. The debriefing subfactor included five items focused on helping PSTs reflect on their lesson through questioning, analysis of student work, or data analysis. The developing practice subfactor involved four items focused on modeling specific instructional skills or providing opportunities to practice outside of regular instruction. The collaborative coaching practice included two items measuring the frequency of co-teaching and co-planning activities. The specific factor about the frequency of coaching in the instructional domain included 11 items aligned with the corresponding domain in the TEAM observation rubric used in Tennessee. All factors again met thresholds for good to excellent psychometric properties (see Ronfeldt et al., 2020).

Survey Response Rates

We calculated response rates for our three main survey instruments: PSTs' pre-survey, PSTs' post-survey, and CMs' survey. The results are reported in Table 3. Across cohorts, our response rate was 61.0% for the PST pre-survey, 40.3% for the PST post-survey, and 56.2% for the CM survey. We find that survey response rates were lower for *Cohort 2* than *Cohort 1* for all three survey instruments (pre-survey −12.1pp, post-survey −9.6pp, and mentor survey −39.2pp). Importantly, for *Cohort 2*, the PST post-survey and many waves of the mentor survey were administered in the spring of 2020 during the first wave of COVID-19, a time when many PSTs, CMs, and students were moving to remote instruction and experiencing substantial turmoil, likely accounting for some of the observed declines in response rates. Comparing response rates across treatment status, we do not find evidence of differential response rates across conditions.

In Appendix Table 1, we compare respondents to non-respondents across all three surveys. Overall, we find that PSTs who responded to the surveys are somewhat different from those that did not, though there is no difference for CMs. The patterns are somewhat expected from what has been reported in the survey literature. On the pre-survey, we observe that women and more academically successful PSTs were slightly more likely to respond than their peers, though our joint significance test finds no overall difference between the two groups. For the post-survey, we do find that the two groups differ significantly, largely due to the significantly higher GPAs of responding PSTs.

Table 3. Survey response rates.

	(1) Pre-survey	(2) Post survey	(3) Mentor survey
Combined	0.610	0.403	0.562
Cohort 1	0.671	0.452	0.761
Cohort 2	0.550	0.356	0.369

Notes. This table reports the response rates on PSTs' pre-placement and post-placement surveys as well as CM's surveys. *t*-tests comparing response rates between treatment and control groups show no evidence of differential response rates by condition.

Pre-Service Teacher Clinical Assessments

New to this replication study, we collected clinical assessment scores, which are observational measures of PSTs' instructional effectiveness during student teaching. PSTs were assessed up to five times during their clinical placements, twice during their first semesters and thrice during their second semesters. CMs completed three of these evaluations (twice during the second semester), and university field supervisors the other two (once per semester).

PSTs were evaluated on a 19-item rubric closely aligned with the state's Tennessee Educator Acceleration Model (TEAM) rubric, used as part of the inservice teacher evaluation system in the state. These items are divided into three teaching domains—instruction, environment, and planning—and intend to provide a well-rounded assessment of teaching practice. We take the average across all 19 items to generate an overall clinical assessment score, which we then standardize within each year.

Pre-Service Teacher Employment Rates and Observation Ratings

Finally, we link PSTs to state administrative datasets with records of employment and evaluation data. From these, we are able to observe PSTs' employment (i.e., being hired as a teacher of record in a public school in Tennessee) and their scores on the TEAM rubric. Each new teacher in Tennessee is observed at least four times during their first years of teaching and, over these observations, receives at least two ratings in each domain on the TEAM rubric (i.e., instruction, environment, planning, and professionalism). We average scores across these observations and domains to calculate an overall observation rating for each school year. Notably, though we have employment information across both cohorts, observation ratings were not available for *Cohort 2* due to a pause on teacher evaluation during the COVID pandemic.

Analytic Strategy

Our replication of Ronfeldt et al. (2020) follows the same analytic strategy. Our preferred model is a linear regression with fixed effects for placement area—requested grade level and/or subject:

$$Y_{ijk} = \beta_0 + \beta_1 \cdot \ Treat_k + \phi_j + \epsilon_{ijk}$$

where Y_{ijk} is the outcome of interest for PST i in placement area j in district/county k, $Treat_k$ is an indicator variable that takes the value of 1 if a PST completed a clinical placement in a county randomized to receive a recommendation list, ϕ_j is a vector of indicator variables for the field placement areas that PSTs are pursuing, and ϵ_{ijk} is the standard error term clustered at the district level. Combined analyses involving both cohorts also control for average differences in outcomes by year.

For analyses involving clinical assessments as outcomes, we modify the specification above, as each PST is observed up to five times. Our preferred approach therefore involves a multilevel linear regression model with observations (time) nested within PSTs:

$$Y_{tijk} = \beta_0 + \beta_1 \cdot \ Treat_k + \lambda_t + \mu_i + \ \phi_j + \epsilon_{tijk}$$

where Y_{tijk} is the clinical assessment score at time t for individual i in placement area j and district/county k, μ_i are random intercepts for each PST, and λ_t is a vector of indicator variables for the ordinal position of the clinical assessment (i.e., second, third, etc. with first as the reference category). All other variables remain the same as specified above. As above, analyses combining *Cohorts 1* and *2* again also control for year-to-year differences.

The Quasi-Experimental "Extension" Study with New Programs

Design

The second part of this paper evaluates recommendation list use with a new group of EPPs in the state (i.e., the extension study). The goal of this study was to evaluate the extent to which program-wide implementation of recommendation lists could increase the overall instructional quality and experience of recruited mentors among a wider variety (in terms of size and geographic location) of EPPs and with minimal oversight or guidance from the TDoE and research partners. Doing so intended to provide insight into whether the use of recommendation lists could be scaled up across a wider variety of EPPs and under more typical circumstances; for this implementation, we did not examine other outcomes (e.g., survey or workforce outcomes) for PSTs or CMs.

The TDoE invited EPPs across the state to apply to participate in this study through a grant program. Participation was conditional on agreeing to use our recommendation lists to inform CM recruitment for the 2020–21 academic year and to share four years of prior CM data with the research team. Ultimately, three EPPs in the state were selected to participate in this study. We note here that the goal of this study was not to recruit a more representative or diverse EPP pool than the two previous RCTs but rather to involve different EPPs to implement the targeted recommendation lists under more natural circumstances. In particular, we were concerned that the original and replication field experiments were completed in somewhat ideal conditions—with an EPP that not only had longstanding relationships with the research/state partners but also was implementing the initiative with their ongoing support and collaboration.

We developed recommendation lists for all requested placements for the three participating EPPs. The development of the recommendation lists followed the same procedures as in the replication study and the original RCT described above. However, the implementation of the lists differed somewhat from the first two field experiments; namely, we requested partner programs to use the recommendation lists to recruit CMs for all PSTs rather than just for those PSTs randomized to treatment.

We note that implementation of this quasi-experimental extension coincided with the COVID-19 pandemic. Importantly, our outcome measure of interest—the evaluation scores of the teachers who were recruited to serve as mentors—come from the prior academic year, thus leaving them largely unaffected by the pandemic, with the exception of the few teachers serving as mentors in our intervention year who would have been observed in mid-March or later of the previous spring. However, the pandemic almost certainly impacted which teachers chose (and perhaps which were recruited by programs) to serve as mentors, thus creating a contemporaneous shock that would bias any estimates of the impact of our intervention. However, it is difficult to predict the sign of

any such bias, given that we do not have any strong hypotheses as to how the pandemic might have impacted teachers' willingness to host PSTs. As a result, we simply point out the pandemic as a note of caution and encourage the replication of our extension study in more typical conditions.

Analytic Strategy

We leverage year-to-year, within-EPP changes in average CM instructional quality to assess how receiving recommendation lists affected CM recruitment. In detail, we use the following 2-level multilevel regression model with teachers nested in EPPs:

$$Y_{ip} = \beta_0 + \beta_1 \cdot \ Treat_p + \pi_p + \epsilon_{ip}$$

where Y_{ip} is the CM evaluation score for teacher i in EPP p. All evaluation scores are standardized within year to have a mean of 0 and standard deviation of 1 to account for statewide, year-to-year variation in teacher evaluation scores unrelated to our treatment, thus de-trending outcome data. $Treat_{yp}$ is an indicator variable that takes the value 1 when program p received recommendation lists during our quasi-experiment and zero otherwise. π_p are random intercepts for each program. Finally, ϵ_{ip} is the residual term. The coefficient of interest is β_1. We interpret this coefficient as the quasi-experimental effect of receiving recommendation lists on the average instructional quality of recruited CMs when compared to historical CM recruitment efforts of the EPP. The identifying assumption in these models is that EPPs would have continued with their business-as-usual CM recruitment strategies in the absence of our intervention. While this assumption is untestable, Figure 2 at least provides no evidence of a clear and positive pre-trend in average CM instructional quality. Moreover, in separate analyses, we ran regression models that included a linear trend for time to further explore the extent to which possible trends in CM scores unrelated to our quasi-experiment could explain our results. All coefficients for this linear trend were not significant and small in magnitude ($ß = 0.011$, $p > .05$ for average CM observation ratings). Additionally, a post-hoc test of significance cannot reject the null hypothesis that the average CM observation ratings for the five CM cohorts recruited before the introduction of our recommendation lists are the same ($F(4, 876) = 1.77$, $p = .133$), underscoring that we do not observe any trend prior to treatment.

The main limitation of this analytic approach is the absence of an untreated group that we could use to evaluate the effects of unobserved confounders on CM instructional quality for the quasi-experimental academic year. This information would have given us the opportunity to use a difference-in-differences model rather than an interrupted time series. However, CM-PST match information is not available for non-treated EPPs, so this analysis is not feasible in this study.

Results

In this section, we begin by describing results from our original and replication MMR field experiments. Afterwards, we describe results from our quasi-experimental extension to three new programs.

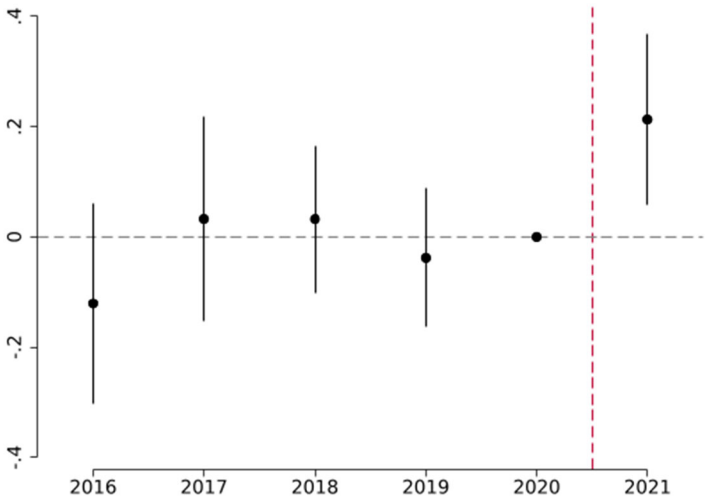

Figure 2. Event study of the effects of receiving a recommendation list on CMs' observation ratings. *Notes*. This figure reports the results of an interrupted time series analysis of the relationship between recruiting clinical mentors using our recommendations lists on CMs' average observation ratings. All outcomes are standardized within each year using all teachers in the state. The dashed red line separates the cohorts of mentors that were recruited using business-as-usual practices (on the left) from those recruited with our recommendation lists (on the right). Each coefficients reports the average CM standardized observation rating of the recruited cohort.

Study 1: Replicating (Cohort 2) the Original (Cohort 1) Field Experiment

Clinical Mentor Contrast between Treatment and Control Districts

Table 4 reports the contrast between treatment conditions on each component of the recommendation index. This table has three panels. Panel A reports the combined (*Cohort 1* and *Cohort 2*) results, while Panels B and C report the results for *Cohorts 1* and *2* broken apart. Across both years of implementation (Panel A), the use of the recommendation lists was able to produce large and statistically significant contrasts between control and treatment CMs in terms of instructional effectiveness and experience. Contrasts were greater than 0.5 standard deviations across all three measures (observation ratings, TVAAS, and experience).[3] Comparing the results for the replication (*Cohort 2*) and original RCT (*Cohort 1*), we find that the experimental contrasts obtained in the replication study appear to be larger in magnitude than those in the original RCT for observation ratings and experience and smaller for TVAAS. However, these differences are not statistically significant. Together, these results suggest that recruiting CMs using our recommendation lists increased the instructional effectiveness and experience of the CM cohort and that these increases are reproducible over the course of multiple years. Notably, as CM recruitment for *Cohort 2* took place in 2019, the results of this portion of our replication were entirely unaffected by the COVID pandemic.

[3]The unadjusted contrast that excludes field fixed effects is similar in magnitude and significance level to the field-effect adjusted estimate: $d = 0.473$, s.e. $= 0.097$, $p < .001$, $N = 145$. Note that we exclude three mentors in singleton requested field placements (i.e., placement subjects/levels for which only one CM was requested) for the main analyses.

Table 4. Contrast on mentor characteristics.

	(1) Mean	(2) Control mean	(3) Treatment mean	(4) Diff	(5) Effect size	(6) *p*-Value
Panel A. Combined						
Observation ratings	0.477	0.307	0.671	0.364	0.557	0.000
TVAAS	0.501	0.327	0.794	0.468	0.505	0.002
Years of experience	15.056	12.792	18.013	5.22	0.578	0.000
Recommendation index	0.458	0.273	0.717	0.443	0.749	0.000
Panel B. Cohort 1						
Observation ratings	0.467	0.317	0.653	0.336	0.477	0.003
TVAAS	0.498	0.304	0.890	0.586	0.698	0.004
Years of experience	15.534	13.274	18.244	4.97	0.517	0.002
Recommendation index	0.470	0.307	0.722	0.415	0.687	0.000
Panel C. Cohort 2						
Observation ratings	0.489	0.297	0.692	0.394	0.654	<0.001
TVAAS	0.504	0.347	0.701	0.354	0.350	0.121
Years of experience	14.512	12.296	17.757	5.461	0.645	<0.001
Recommendation index	0.445	0.238	0.711	0.473	0.808	<0.001

Notes. Joint chi square test for Panel A: χ^2 (4) = 44.02, $p < 0.001$; Panel B: χ^2 (4) = 23.64, $p < .001$; Panel C: χ^2 (4) = 29.77, $p < .001$. $N = 359$, 156 treatment and 144 control mentors.

Interaction Between the Replication Study and Original RCT. Table 5 summarizes descriptive results from exploratory analyses of how the replication study and the original RCTmay have interacted with each other. Since the same set of districts were randomized to receive the recommendation lists in two consecutive years, we can explore how receiving the recommendation lists in the original RCT may have affected CM recruitment in the replication study.[4] We report on two sets of outcomes: overall contrast on the recommendation index and contrast on CM percentiles on the teacher distribution within a placement block (endorsement subject/level-by-county). While the recommendation index is in standard deviation units and can help with comparison with other work in this space, the percentiles offer a more concrete contextualization of our results to the actual distribution of instructional quality and experience.

Reading Table 5 from top to bottom, we find that districts that were randomized to receive recommendation lists in the original RCT but not in the replication study (i.e., Treatment/Control) do not appear to recruit more instructionally effective and experienced CMs (during the replication study) than districts that never received recommendation lists (i.e., Control/Control districts; the reference category). These findings suggest that, since they no longer received recommendation lists in the replication study, these districts mostly reverted to business-as-usual approaches during the replication year and recruited similar mentors as districts that never received recommendation lists.

Districts that move from control to treatment between the two implementation years appear to benefit the most from receiving the recommendation lists. We observe a difference between this group and the Control/Control group of 0.896 standard deviation units on the recommendation index or 31.895 percentiles on the teacher distribution ($p < .001$).

Similarly, districts that received the recommendation lists two implementation years in a row (i.e., Treatment/Treatment) continued to benefit from the recommendation lists (during the replication study). We observe a difference between this group and the

[4]Note that we did not block randomization for receiving recommendation lists in the replication study using treatment status from the original RCT. As a result, these analyses are exploratory in nature only, and we cannot interpret these results causally.

Table 5. Interaction between implementation years.

	(1) Recommendation index	(2) Recommendation percentile
Treatment/Control	0.142	7.981
	(0.144)	(6.073)
Control/Treatment	0.896***	31.895***
	(0.187)	(6.508)
Treatment/Treatment	0.558***	25.003***
	(0.136)	(5.215)
Observations	119	119
R^2	0.297	0.300
Adjusted R^2	0.210	0.213

Notes. Coefficients report differences from districts that never received recommendation lists (Control/Control). Standard errors in parentheses. ***$p < .001$.

Control/Control group of 0.558 standard deviation units on the recommendation index or 25.003 percentiles on the teacher distribution ($p < .001$). This finding is consistent with an explanation that repeated use of the recommendation lists did not fully exhaust the untapped pool of instructionally effective and experienced teachers to serve as mentors; however, the somewhat tempered contrast may indicate slightly diminishing returns over time, though a post-hoc test shows that the differences between estimates for the Control/Treatment and Treatment/Treatment groups are not statistically significant ($d = 0.337$, $F(1, 18) = 3.44$, $p = .080$ for recommendation index and $d = 6.89$, $F(1, 18) = 1.33$, $p = .265$ for percentiles).

Treatment Effects on Pre-Service Teacher Self-Reported Survey Outcomes

Table 6 reports estimates of treatment effects on PST self-reported survey outcomes, comparing PSTs who completed their clinical experiences in districts that were randomized to receive recommendation lists with PSTs who completed their clinical experiences in districts that recruited CMs following their business-as-usual recruitment strategies.

We report three survey-based outcomes categories—feelings of readiness, mentoring frequency and mentoring satisfaction—each with its own panel in Table 6. We report the estimates combining *Cohorts 1* and *2* in Column 1, the original RCT (*Cohort 1*) in column 2, and the replication RCT (*Cohort 2*) in column 3. All estimates include fixed effects for clinical placement area (requested grade level/subject) and cluster standard errors at the district level.

We notice two main takeaways. First, across years, PSTs who completed their clinical placements in treated districts reported feeling more ready to teach than their peers in control districts. These results are consistent for the overall feelings of readiness factor and its twosub-factors, readiness in questioning skills and readiness in other instructional skills. Compared to the original RCT, though, these differences are smaller in magnitude and not statistically significant in the replication study. However, we again point out that survey data collection for the replication study happened during April and May 2020, which coincided with the first month of online instruction resulting from the school closures due to COVID-19. It is likely, though not the only possible explanation, that PSTs' feelings of readiness might be more sensitive to these changes in

Table 6. Treatment effects on self-reported pre-service teacher surveys.

	(1) Combined	(2) Cohort 1	(3) Cohort 2
Panel A: Feelings of readiness			
Feeling of readiness – teaching skills	0.463*	0.593**	0.250
	(0.198)	(0.226)	(0.227)
Readiness in questioning skills	0.478*	0.637**	0.235
	(0.186)	(0.230)	(0.217)
Readiness in other instructional skills	0.448*	0.548*	0.266
	(0.216)	(0.225)	(0.280)
Panel B: Mentoring frequency			
Mentoring frequency	0.148	0.181	0.100
	(0.125)	(0.147)	(0.198)
Common mentoring practices	0.172	0.143	0.187
	(0.139)	(0.184)	(0.192)
Data-driven mentoring practices	0.213	0.236	0.197
	(0.153)	(0.201)	(0.202)
Collaborative coaching practices	0.107	0.205+	−0.063
	(0.120)	(0.111)	(0.275)
Modeling coaching practices	0.101	0.141	0.080
	(0.141)	(0.186)	(0.206)
Panel C: Mentoring satisfaction			
Coaching satisfaction	−0.106	−0.143	0.018
	(0.121)	(0.171)	(0.193)
Support and feedback	−0.131	−0.181	−0.017
	(0.126)	(0.170)	(0.207)
Autonomy and encouragement	−0.081	−0.105	0.052
	(0.121)	(0.179)	(0.188)

Notes. This table reports treatment effects on outcomes from the PSTs' post-survey. All regressions include field placements fixed effects. Standard errors clustered at the district by subject block in parentheses. $+p < .10$; $*p < .05$; $**p < .01$.

teaching format and the general effects of COVID-19 related life changes than to treatment, perhaps overriding any effects of the lists themselves.

Second, we find no significant differences between conditions in terms of PSTs' reports about the amount or quality of mentoring they received. Overall and in each year of implementation, there were no significant differences between control and treatment PSTs in terms of the frequency of mentoring (overall, common, data-driven, and collaborative) they experienced. There were also no differences in PSTs' overall satisfaction with the coaching they received, their assessment of the support and feedback offered by their CMs, and the balance of autonomy and support their CMs provided. Though all coefficients are non-significant, those on mentoring frequency measures collectively tended to trend positive, while those on coaching satisfaction/quality measures tended to trend negative.

Treatment Effects on Clinical Mentor Self-Reported Survey Outcomes

Table 7 reports the contrast in self-reported survey outcomes between CMs in districts that were randomized to receive recommendation lists and CMs in districts following business-as-usual recruitment strategies. The columns in this table follow the same structure as the one described above for Table 6.

Across years and measures, there were no significant differences between control and treatment CMs in terms of the frequency of mentoring practices they reported offering their PSTs. Treatment CMs tend to report somewhat more frequent mentoring, but differences are non-significant.

Table 7. Treatment effects on clinical mentor self-reported survey outcomes.

	(1) Combined	(2) Cohort 1	(3) Cohort 2
Frequency of mentoring practices	0.111	0.058	0.231
	(0.186)	(0.207)	(0.326)
Debriefing	0.276	0.283	0.267
	(0.195)	(0.220)	(0.343)
Developing practice	−0.010	−0.101	0.244
	(0.149)	(0.155)	(0.308)
Collaborative coaching practices	0.075	0.002	0.181
	(0.194)	(0.215)	(0.360)
Coaching frequency in instruction domain	0.204+	0.190	0.177
	(0.117)	(0.149)	(0.390)

Notes. This table reports the treatment effects on CMs' mentoring survey. All regressions include field placements fixed effects. Standard errors clustered at the district level in parentheses. $+p < .10$.

As mentioned in the methods section, we observe some differences between PSTs and CMs who responded and did not respond to our surveys. As a robustness check, we included PST and CM characteristics in our models. We find qualitatively similar results between the models with and without these control variables. These results are available upon request.

Treatment Effects on Clinical Assessment Scores

In Table 8, we report treatment effects of the initiative on PSTs' instructional effectiveness as measured by the clinical assessment ratings completed by their field supervisors and CMs during student teaching. The columns follow the same progression as in prior tables.

Beginning with the combined results (across *Cohorts 1* and *2*), we find that PSTs assigned to treatment CMs were rated as significantly more instructionally effective than PSTs assigned to control CMs. Estimates were nearly identical across all three domains of teaching captured on the clinical assessment rubric. Unlike for survey outcomes, though, differences were greater for *Cohort 2* than for *Cohort 1*; in fact, coefficients on the latter are non-significant.

When we separate the treatment effects between field supervisor and CM ratings (see Appendix Table 2), we notice that the treatment effect is larger in magnitude and only statistically significant for CMs, though the non-significant coefficients for field supervisors are still positive. It is worth noting that our estimate for field supervisor ratings is somewhat less precise for *Cohort 2* than for *Cohort 1*, as clinical placements were cut short due to COVID-19, making it impossible for nearly all field supervisors to complete their second set of clinical evaluations and substantially reducing the number of observations.

Treatment Effects on Employment Rates and Observation Ratings

Finally, we follow both cohorts of PSTs into the workforce to observe their employment rates and TEAM observation ratings. It is worth noting here that *Cohort 1* entered the job market in fall 2019, while *Cohort 2* entered the job market in fall 2020. Since Tennessee suspended its teacher evaluation system for the 2020–2021 school year in response to the COVID-19 pandemic, we only have observation ratings for the first cohort.

Table 9 reports the contrast in the employment rates and instructional effectiveness—as measured by first-year observation ratings—of treatment and control PSTs. There were no significant differences between PSTs who had been mentored by treatment

Table 8. Treatment effects on clinical assessment scores.

	(1) Combined	(2) Cohort 1	(3) Cohort 2
Treatment	0.216**	0.171	0.319**
	(0.083)	(0.106)	(0.122)
Constant	−0.379	−0.127	−0.458
	(0.323)	(0.605)	(0.361)
Observations	1157	676	481

Notes. This table reports treatment effects on PSTs' clinical assessment scores. All models are multilevel mixed effects models that include random effects for PST, field placement fixed effects, and controls for observation ordinal position. Standard errors in parentheses. $^{**}p < .01$.

CMs (i.e., those recruited using the recommendation lists) and those mentored by control CMs (i.e., those recruited using business-as-usual recruitment practices) on either employment rates or observation ratings. Though not statistically significant, outcomes consistently trended positive across measures, suggesting employment rates and observation ratings may have been slightly stronger for treatment graduates.[5]

Study 2: Quasi-Experimental Extension to New Programs

We report the results of our quasi-experimental extension with three new EPPs in Table 10. In this table, the constant represents the average CM observation rating in these EPPs during the pretreatment period, while the list use coefficients represent the differences in CM observation ratings for the academic year during which CMs were recruited using the recommendation lists.

During the year in which partner programs' leadership received recommendation lists, they recruited CMs whose observation ratings were significantly greater by 0.226 standard deviations ($p < .01$) than in prior years when they did not have access to recommendation lists. Scanning the other four columns of this table, we observe some heterogeneity in the point estimate for specific subdomains of the teacher observation rubric, with the biggest effect concentrated in CMs' planning domain scores ($ß = 0.244$, $p < .01$) and the smallest effect on CMs' environment domain scores ($ß = 0.101$, $p < .10$). Notably, effect sizes are similar in magnitude to the experimental results that we observed for the implementation of the first two cohorts of the initiative. This could suggest that the effects of having access to recommendation lists are stable across kinds of EPPs and do not diminish when EPPs receive minimal oversight or guidance regarding how to use the lists.

Discussion

This paper reports on a series of studies meant to improve the instructional effectiveness and experience of teachers who serve as CMs by using historical administrative data to guide recruitment. We show through the original (*Cohort 1*) and reproduction (*Cohort 2*) field experiments with the same program (TTU) that partner districts randomly

[5]As only about 80% of the PSTs in *Cohort 1* ended up employed in Tennessee, we have even more reduced power to precisely estimate the impacts of the initiative on observation ratings.

Table 9. Treatment effects on workforce outcomes.

	(1)	(2)	(3)	(4)
	Employment			
	Combined	Cohort 1	Cohort 2	Observation ratings
Treatment	0.066	0.045	0.100	0.096
	(0.046)	(0.058)	(0.070)	(0.144)
Constant	0.768***	0.786***	0.644***	3.530***
	(0.037)	(0.045)	(0.043)	(0.092)
Field fixed effect	Yes	No	No	Yes
Year fixed effect	Yes	No	No	No
Observations	299	152	144	71
R^2	0.162	0.099	0.224	0.102
Adjusted R^2	0.108	0.014	0.152	−0.031

Notes. This table reports the effects of receiving the recommendation lists on PSTs' employment and teacher evaluation outcomes. All models are multilevel mixed effects models that include random effects for educator preparation programs. Clustered standard errors at the district level in parentheses. ***$p < .001$.

Table 10. Interrupted time series results on clinical mentor instructional quality.

	(1)	(2)	(3)	(4)
	OR	Instr.	Env	Plan
Treatment	0.226***	0.233**	0.101	0.244**
	(0.068)	(0.075)	(0.068)	(0.082)
Constant	0.441***	0.442***	0.419***	0.344***
	(0.075)	(0.068)	(0.073)	(0.038)
Observations	1009	1001	987	994

Notes. This table reports the effects of receiving the recommendation lists on CMs' average observation ratings and subdomains of the teacher observation rubric. All models are multilevel mixed effects models that include random effects for educator preparation programs. Clustered standard errors at the district level in parentheses. **$p < .01$; ***$p < .001$.

assigned to receive the recommendation lists were able to recruit CMs who had significantly and meaningfully greater observation ratings, TVAAS scores, and years of experience (by 0.5 standard deviation units across measures) than districts using business-as-usual recruitment practices. Though field experiments often fail to replicate, this initiative successfully produced a treatment contrast of similar magnitudes across years of implementation. Even districts that received the recommendation lists two years in a row were able to recruit substantially more experienced and instructionally effective CMs for *Cohort 2*, suggesting that the pool of experienced and effective teachers is large enough to implement the initiative successfully in consecutive years. Districts that previously but no longer received lists appeared to mostly revert to business-as-usual practices, implying the importance of continued access to a ranking of recommended mentors.

Meanwhile, results from the quasi-experimental extension study indicate that these effects can translate beyond the initial partner program, illustrating that the use of recommendation lists has the potential to increase the instructional effectiveness of teachers serving as CMs across a variety of types of programs. Importantly, this implementation was carried out with little guidance or oversight on the part of the research team or partners at the TDoE, suggesting that merely providing recommendation lists, even without much management or support, can increase the quality of the CM pool. It was also carried out during the pandemic, providing both a note of caution about validity and a possible testament to the robustness of our intervention to unpredictable external shocks. These results illustrate that scaled-up use of recommendation lists across many

programs is a promising low-cost, low-impact approach to improving clinical preparation across the state.

The original and replication field experiments with TTU also demonstrate that increasing the instructional effectiveness and experience of CMs has positive downstream effects on the PSTs they mentor. In particular, PSTs who were mentored by CMs reported feeling better prepared to teach and received higher clinical assessment scores during student teaching, which suggest that the intervention had a positive impact on their observed instructional effectiveness during student teaching. These results are not only consistent with effects found by Ronfeldt et al. (2018) and Goldhaber et al. (2022) but also more clearly demonstrate that these positive effects are driven solely by CM characteristics, rather than in tandem with field placement school characteristics. In other words, these results are the strongest to date that learning to teach with more instructionally effective and experienced CMs causes PSTs to be more instructionally effective themselves. These findings add to the large and still growing body of evidence that support policies, like those in Tennessee, that set minimum requirements in levels of instructional effectiveness for teachers to serve as CMs.

Notably, the positive effects on clinical assessment scores are mostly driven by mentor teacher ratings, while we do not observe a consistent treatment effect on field supervisor ratings. The fact that CM ratings appear to drive more of these treatment effects may raise some concern since CM recruitment is central to the intervention. In particular, there are two concerns we must consider. First, treatment CMs could have become aware of their treatment status and, as a result, rated more leniently. Second, more instructionally effective CMs (i.e., those in treatment) may simply be more lenient raters in general. We believe that neither of these explanations is likely. As to the former, we designed the intervention such that only recruiters were aware of the intervention, requesting that they ensure CMs remain blind to their condition and to the purpose of the intervention, which, according to all correspondences, was in fact the case. Regarding the latter, the only existing study on this topic suggests the opposite to be true—namely, that more instructionally effective CMs are actually harsher raters (Goldhaber et al., 2022). All these explanations, though, are speculative and in need of further study.

Analyses of employment and teacher observation ratings, though limited by sample constraints in part brought on by the pandemic (i.e., the absence of observation ratings for *Cohort 2*), are consistent directionally with our hypotheses and with clinical assessment results. Though differences were small in magnitude and not statistically significant, we find that PSTs in the treatment condition were slightly more likely to be employed in the state and to receive higher average observation ratings. Future research should reproduce this initiative across more programs and years to achieve adequate statistical power and test whether observed trends represent real effects or not.

In the quasi-experimental extension study with three new programs, we cannot confirm whether these increases in the quality of the CM pool among scale-up EPPs are substantial enough to translate into impacts on PST instructional effectiveness. However, prior work (Ronfeldt et al., 2018a) found that a standard-deviation increase in CM observation ratings correlates with an increase of between 0.10 and 0.15 standard deviations in PST observation ratings. At this rate, we would expect placement with a CM recruited under the use of the recommendation lists to produce an increase of

between 0.02 and 0.03 standard deviation units in PST ORs (roughly equivalent to between 15 and 20% of the growth observed during the first year of teaching). However, future work should test this observation more formally rather than relying on back-of-the-envelope calculations.

While the evidence points to consistently positive impacts on the instructional effectiveness and experience of CMs and likely positive impacts on the outcomes of their PSTs, a final and critical direction of future inquiry involves how this intervention may have had unintended consequences on other less directly implicated CM and PST characteristics. One such potential repercussion involves how the use of recommendation lists might impact the racial and gender diversity of the CM pool. Teachers of color and men make up less than one-fourth of the teaching workforce generally (Carver-Thomas, 2018) and an even a smaller share of the CM pool specifically (Ronfeldt et al., 2018a). Since the MMR algorithm depends upon teachers' observation ratings, which tend to be lower for teacher of color and for men (Campbell, 2020; Grissom & Bartanen, 2022), it is logical to worry that its implementation could result in the over-recruitment of White women to serve as CMs.

Our studies were not explicitly designed to identify the generalizable effects of this intervention on CM diversity, especially given the racial and gender homogeneity of CMs at our partner programs. However, we are able to initially explore how the demographics of the CM pool differed between districts (or programs) with lists and those using business as usual practices. In Appendix Table 3, we consider the effects of the intervention on CM race and gender using our same analytic strategies; columns (1) – (3) explore impacts at our primary partner program in both implementation years, while column (4) investigates effects for the three new programs in our extension study. Across columns, we find no significant effects of the intervention on CM race. Effects on gender are less consistent; the intervention appears to increase the proportion of male CMs for *Cohort 1* (column 2) but has the opposite effect (i.e., more female CMs) for the extension programs (column 4). Overall, the effects of the intervention on racial and gender diversity are unclear among our partner programs, and we caution against interpreting these preliminary analyses as evidence for the presence or absence of any biases in the algorithmic development of recommendation lists.[6] More research is needed to further interrogate the impacts of this intervention (or similar ones) on CM racial and gender diversity given the importance of diversifying our teaching workforce and those who mentor them.

Despite their limitations, including the unforeseeable impact of the COVID-19 pandemic, our studies offer examples of the affordances of conducting rigorous experimental research in teacher education, including efforts to replicate and scale up the initial promising results of an RCT. Following best practice recommendations (Makel & Plucker, 2014;

[6]Although there are no observed effects on CM race using our preferred models, preliminary analyses using *t*-tests (i.e., not regression adjusted) revealed significant effects for the extension study involving three new programs. Specifically, 82 percent of CMs were White prior to the intervention, as compared with 90 percent following the intervention. This was driven largely by one program which, prior to the intervention, had a CM pool that was the most racially diverse, recruiting 49 percent CMs of color on average. However, in the year following the intervention (2020–21), all twelve CMs were White. It is important to underscore, though, that this same program had a dramatic decline in cohort size from 2020 ($N = 70$) to 2021, and that the other two programs, in which enrollment was more stable, saw no significant impacts on CM race, raising questions about whether the impact at this one program was related to the pandemic.

Steiner et al., 2019; Wong & Steiner, 2018), we designed and implemented the two studies reported in this paper to build upon the initial RCT. This allowed us to integrate the replication and scale-up efforts into the initial planning with our research partners and participants, thus reducing the research and implementation effort that a separate replication study would have required; at the same time, our replication claims were strengthened by a new quasi-experimental extension involving different EPPs in new labor markets. We believe both studies provide illustrations of the promise of designing and executing field experiments across multiple implementation years and settings in order to assess the stability and generalizability of positive early findings.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Open Research Statements

Study and Analysis Plan Registration

There is no study and analysis plan registration associated with this manuscript.

Data, Code, and Materials Transparency

The data, code, and materials underlying the results reported in this manuscript are not publicly available.

Design and Analysis Reporting Guidelines

The submitted manuscript was not accompanied by a completed JREE Randomized Trial Checklist.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Replication Statement

This manuscript reports a replication of the following study: Ronfeldt, M., Bardelli, E., Mullman, H., Truwit, M., Schaaf, K., Baker, J. (2020). Improving Student Teachers' Readiness to Teach Through Recruitment of Instructionally Effective and Experienced Cooperating Teachers: A Randomized Experiment. Educational Evaluation and Policy Analysis, 42(4), 551–575.

Funding

This work was supported by the Institute of Education Sciences (PR/Award R372A150015) and the Institute of Education Sciences (PR/Award R305B150012).

References

Bastian, K. C., Patterson, K. M., & Carpenter, D. (2020). Placed for success: Which teachers benefit from high-quality student teaching placements? *Educational Policy*, *36*(7), 1583–1611. https://doi.org/10.1177/0895904820951126

Campbell, S. L. (2020). Ratings in black and white: A quantcrit examination of race and gender in teacher evaluation reform. *Race Ethnicity and Education*, *26*(7), 815–833. https://doi.org/10.1080/13613324.2020.1842345

Carver-Thomas, D. (2018). *Diversifying the teaching profession: How to recruit and retain teachers of color*. Learning Policy Institute.

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

Goldhaber, D., Krieg, J., Theobald, R., & Goggins, M. (2021). Front end to back end: Teacher preparation, workforce entry, and attrition. *Journal of Teacher Education*, *73*(3), 253–270. https://doi.org/10/gk45d4

Goldhaber, D., Ronfeldt, M., Cowan, J., Gratz, T., Bardelli, E., & Truwit, M. (2022). Room for improvement? Mentor teachers and the evolution of teacher preservice clinical evaluations. *American Educational Research Journal*, *59*(5), 1011–1048. https://doi.org/10/gpfg8q

Grissom, J. A., & Bartanen, B. (2022). Potential race and gender biases in high-stakes teacher observations. *Journal of Policy Analysis and Management*, *41*(1), 131–161. https://doi.org/10.1002/pam.22352

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10/gcmc76

Matsko, K. K., Ronfeldt, M., & Nolan, H. G. (2022). How different are they? Comparing teacher preparation offered by traditional, alternative, and residency pathways. *Journal of Teacher Education*, *73*(3), 225–239. https://doi.org/10/gmkbcn

Ronfeldt, M. (2021). *Links among teacher preparation, retention, and teaching effectiveness.* National Academy of Education Committee on Evaluating and Improving Teacher Preparation Programs. National Academy of Education. https://files.eric.ed.gov/fulltext/ED615304.pdf.

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

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018a). 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. (2018b). *Identifying promising clinical placements using administrative data: Preliminary results from ISTI placement initiative pilot* (Working Paper No. 189). National Center for Analysis of Longitudinal Data in Educational Research. https://caldercenter.org/publications/identifying-promising-clinical-placements-using- administrative-data-preliminary-results

Ronfeldt, M., Reininger, M., & Kwok, A. (2013). Recruitment or preparation? Investigating the effects of teacher characteristics and student teaching. *Journal of Teacher Education*, *64*(4), 319–337. https://doi.org/10.1177/0022487113488143

Ronfeldt, M., Matsko, K. K., Nolan, H. G., & Reininger, M. (2021). Three different measures of graduates' instructional readiness and the features of preservice preparation that predict them. *Journal of Teacher Education*, *72*(1), 56–71. https://doi.org/10.1177/0022487120919753

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie*, *227*(4), 280–292. https://doi.org/10/ggxrw6

Simpson, A. (2022). A recipe for disappointment: Policy, effect size, and the winner's curse. *Journal of Research on Educational Effectiveness*, *16*(4), 643–662. https://doi.org/10.1080/19345747.2022.2066588

Wong, V. C., & Steiner, P. M. (2018). *Replication designs for causal inference (No. 62).* University of Virginia. http://curry.virginia.edu/uploads/epw/62_Replication_Designs.pdf

Appendix A

Table A1. Comparison of survey respondents and non-respondents.

	(1) mean	(2) non-resp	(3) Resp	(4) Diff	(5) E.S.	(6) p-value	(7) *N* Non-Resp	(8) *N* Resp
Panel A. PST Pre-Survey								
Woman	0.868	0.821	0.898	0.078	0.239	0.047	114	176
White	0.975	0.958	0.986	0.028	0.192	0.111	114	176
Current GPA	3.560	3.512	3.591	0.080	0.230	0.056	114	176
ACT	22.81	22.72	22.86	0.141	0.053	0.661	111	171
Praxis	168.7	167.1	169.5	2.475	0.260	0.054	86	156
Panel B. PST post-survey								
Woman	0.868	0.846	0.902	0.056	0.171	0.154	176	114
White	0.975	0.965	0.990	0.025	0.168	0.161	176	114
Current GPA	3.560	3.511	3.636	0.125	0.365	0.003	176	114
ACT	22.81	22.76	22.87	0.103	0.039	0.748	171	111
Praxis	168.7	168.0	169.5	1.499	0.157	0.230	141	101
Panel C. Mentor survey								
Rec Index	0.504	0.473	0.526	0.053	0.084	0.472	124	176
OR (Std)	0.496	0.488	0.502	0.015	0.022	0.854	124	176
TVAAS	0.576	0.503	0.622	0.120	0.125	0.450	59	93
Experience (Std)	0.442	0.396	0.474	0.078	0.075	0.522	124	176

Notes. Effect sizes are calculated as the "covariate-adjusted mean difference divided by the unadjusted pooled within-group SD" (What Works Clearinghouse, 2017, p. E-4). Panel A χ^2 (5) = 8.724, p = .121; Panel B χ^2 (5) = 11.874, p = .037; Panel C χ^2 (4) = 0.912, p = .923.

Table A2. Treatment effects on clinical assessment scores by rater.

	(1) Combined	(2) Cohort 1	(3) Cohort 1	(4) Cohort 2	(5) Cohort 2
		Supervisors	Mentors	Supervisors	Mentors
Treatment	0.216**	0.114	0.265+	0.185	0.421**
	(0.083)	(0.117)	(0.148)	(0.139)	(0.149)
Constant	−0.379	−0.190	0.127	0.010	−0.856+
	(0.323)	(0.665)	(0.836)	(0.399)	(0.452)
Observations	1157	408	268	202	279

Notes. This table reports treatment effects on PSTs' clinical assessment scores. All models are multilevel mixed effects models that include random effects for PST, field placement fixed effects, and controls for observation ordinal position. Standard errors in parentheses. +$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$.

Table A3. Treatment effects on CM race and gender.

	(1)	(2) Primary partner program ("replication")	(3)	(4)
	Combined	Cohort 1	Cohort 2	"Extension"
Panel A. Proportion of white mentors				
Treatment	0.010	0.022	0.000	0.030
	(0.010)	(0.021)	(.)	(0.031)
Constant	0.971***	0.962***	1.000	0.792***
	(0.012)	(0.017)	(.)	(0.104)
Observations	298	152	143	1086
Panel B. Proportion of women mentors				
Treatment	−0.022	−0.093***	0.052	0.055
	(0.027)	(0.024)	(0.043)	(0.034)
Constant	0.880***	0.917***	0.849***	0.809***
	(0.027)	(0.009)	(0.038)	(0.013)
Observations	299	152	144	1086
Field Fixed Effect	Yes	Yes	Yes	No
Year Fixed Effect	Yes	No	No	No
Linear Time Trend	No	No	No	No

Notes. This table reports the effects of receiving the recommendation lists on CTs' race and gender. Models involving our primary partner program are OLS regressions with fixed effects as noted above, while analyses of the extension programs involve multilevel mixed effects models that include random effects for educator preparation programs. Clustered standard errors at the district level in parentheses. ***$p < .001$.