American Educational Research Journal
October 2022, Vol. 59, No. 5, pp. 1011–1048
DOI: 10.3102/00028312211066867
Article reuse guidelines: sagepub.com/journals-permissions

Room for Improvement? Mentor Teachers and the Evolution of Teacher Preservice Clinical Evaluations

Dan Goldhaber
American Institutes for Research
University of Washington
Matt Ronfeldt
University of Michigan
James Cowan
American Institutes for Research
Trevor Gratz
University of Washington
Emanuele Bardelli
Matt Truwit
University of Michigan

The clinical teaching experience is one of the most important components of teacher preparation. Prior observational research has found that more effective mentors and schools with better professional climates are associated with better preparation for teacher candidates. We test these findings using an experimental assignment of teacher candidates to placement sites in two states. Candidates who were randomly assigned to higher quality placement sites experienced larger improvements in performance over the course of the clinical experience, as evaluated by field instructors (a.k.a university instructors). The findings suggest that improving clinical placement procedures can improve the teaching quality of candidates.

Keywords: clinical education, student teaching, teacher education/development, experimental research

Introduction

Teacher preparation has emerged as a high profile area of focus for reforms of policy and practice (American Association of Colleges of Teacher Education [AACTE], 2010). The clinical practice that teacher candidates complete as a requirement for teaching in most states is seen as foundational to their development (AACTE, 2018, National Council for Accreditation

of Teacher Education, 2010); indeed, it has been identified as "a key component—even 'the most important' component—of preservice teacher preparation" (Anderson & Stillman, 2013, p. 3). Yet it is this aspect of teacher preparation that has been judged by some to be lacking, particularly with regards to ensuring that teacher candidates receive high quality mentoring during their clinical placements (also known as student teaching or internship placements; AACTE, 2010; National Council for Accreditation of Teacher Education, 2010).

The role of clinical placements is potentially profound as the guidance and evaluations provided by mentor teachers to teacher candidates is usually the first official counseling and feedback that teacher candidates receive in an authentic classroom setting.[1] An emerging body of observational research documents associations between the quality of a candidate's clinical placement, as measured by mentor effectiveness or school quality, and their future performance in the classroom (Goldhaber, Krieg, Theobald, 2020; Ronfeldt, 2015; Ronfeldt, Bardelli, et al., 2020; Ronfeldt, Brockman, & Campbell, 2018; Ronfeldt, Matsko, et al., 2018; Ronfeldt, Matsko, et al., 2020).

Dan Goldhaber is the Director of the Center for Education Data & Research (CEDR, cedr.us) at the University of Washington and the Director of the Center for Analysis of Longitudinal Data in Education Research (CALDER, caldercenter.org) at the American Institutes for Research, 3876 Bridge Way North, Suite 201, Seattle, WA 98103, USA; email: *dgoldhaber@air.org*. His work focuses on issues of educational productivity and reform at the K–12 level, including the broad array of human capital policies that influence the composition, distribution, and quality of teachers in the workforce, and connections between students' K–12 experiences and postsecondary outcomes.

Matt Ronfeldt is an associate professor of educational studies and a core faculty member of the Institute of Education Sciences Predoctoral Training Program in Causal Inference in Educational Policy Research at the University of Michigan. His scholarship focuses on how to improve teaching quality, particularly in schools and districts serving marginalized communities.

James Cowan is a researcher at the Center for Analysis of Longitudinal Data in Education Research focused on teacher labor markets, teaching effectiveness, and education policy.

Trevor Gratz is a researcher at the Center for Education Data and Research at the University of Washington. His research centers on how education policy shapes student outcomes, and in particular focuses on the intersection of health policy and educational attainment.

Emanuele Bardelli is a doctoral candidate in educational studies and a fellow of the Institute of Education Sciences Predoctoral Training Program in Causal Inference in Educational Policy Research at the University of Michigan. His research interests include teacher learning and supports for instructional improvement, with emphasis on early career teachers.

Matt Truwit is a doctoral candidate in quantitative research methods in education and a master's student in statistics at the University of Michigan. His research explores the influences of school and state policies and programs on educational (in)equity.

Importantly, however, there is little *causal* evidence about the ways that mentors or clinical placements influence teacher candidates.

In this study, we draw on data from within a teacher education program (TEP) randomized control trial to assess the plausibly causal relationship between the quality of clinical placements and the growth of measurable teaching skills. As part of the experiment, we randomly assigned candidates to two lists of potential placements that differed in measures of teacher and school quality. At different points throughout the clinical placement, university field instructors and mentor teachers each provided up to three assessments of candidates' teaching practice using clinical evaluation rubrics aligned with inservice teacher evaluation frameworks. Prior research has found both that such preservice evaluations can predict inservice effectiveness (Chen et al., 2019) and that there is considerable heterogeneity in the rate of improvement among teacher candidates during their clinical placements (Vagi et al., 2019). We exploit the randomization of teacher candidates to either higher or lower quality placements to study the effects of clinical placement quality on the growth of clinical evaluation ratings provided by university faculty, henceforth referred to as field instructors.

We find that candidates randomly assigned to "higher quality" (we describe our definitions of quality in the Background on Mentor Teachers and Clinical Evaluations section) clinical placements improved their clinical practice by 0.06 standard deviations more than candidates randomly assigned to "lower quality" placements. We use evaluations by university field instructors conducted at several points during the intervention as our measure of clinical practice. Prior research has shown that preservice clinical evaluations are predictive of teacher performance in the classroom, although there is some evidence that they can be influenced by classroom assignments (Bastian et al., 2018; Campbell & Ronfeldt, 2018; Chen et al., 2019, Goldhaber, Cowan, & Theobald, 2017; Steinberg & Garrett, 2016). We therefore primarily focus on the growth in candidates' clinical evaluation ratings, rather than their overall ratings, to help account for the subjective nature of these evaluations. As we describe in the Methods section, growth models control for time invariant influences of classroom context, allowing for the recovery of a plausibly causal growth treatment effect. Although nonexperimental estimates of the contributions of individual school and mentor characteristics are imprecise, they suggest that the teaching experience and observational ratings of mentor teachers positively influence candidates' observed performance. These findings on teacher candidates are consistent with prior observational studies of inservice teacher outcomes, suggesting that candidates learn more measurable teaching skills from higher quality clinical placements (Goldhaber, Krieg, & Theobald, 2020; Ronfeldt, Bardelli, et al., 2020; Ronfeldt, Brockman, & Campbell, 2018; Ronfeldt, Matsko, et al., 2018; Ronfeldt, Matsko, et al., 2020).

Background on Clinical Practice and the Improving Student Teaching Initiative

Background on Mentor Teachers and Clinical Evaluations

Clinical placements are supervised by two kinds of educators: university-based field instructors who are employees of the university where the program is based, and mentor/cooperating teachers who are teachers employed by the local schools/districts in which clinical practice takes place (Cuenca et al., 2011). Field instructors (also referred to as "university supervisors") are theorized to help teacher candidates bridge and apply the theory they learn in the didactic portion of their education to their hands-on clinical education (Cuenca et al., 2011). Field instructors also offer "an objective, third-party viewpoint in the student teaching triad" (Gareis & Grant, 2014), and they provide formalized assessments of teacher candidates' skills in the form of clinical evaluations.

Numerous qualitative studies (Clarke et al., 2014; Ganser, 2002; Graham, 2006; Hoffman et al., 2015; Zeichner, 2009) document the myriad roles mentors play in the development of teacher candidates: mentor teachers provide concrete examples of classroom preparation, instructional leadership, and student engagement; and they help induct teacher candidates into school practices and processes. Perhaps most importantly, mentors provide teacher candidates with feedback about their student teaching, often also via formalized clinical evaluations.

Precisely how teachers serving as mentor influence mentees is not well understood, but feedback about student teaching is thought to play a role (Glenn, 2006). Surveys of teacher candidates suggest that teacher candidates value feedback from their mentors (Connor & Killmer, 1995). There is some quantitative evidence that broadly buttresses the importance of feedback. Boyd et al. (2009) find that teachers whose clinical placement experiences incorporate significant oversight and feedback, and are congruent with their assignments as first-year teachers, tend to be more effective teachers. Matsko et al. (2018) find that teacher candidates who reported more frequent and higher quality feedback from their mentor teachers felt better prepared to teach at the end of preparation. And, Ronfeldt, Matsko, et al. (2020) find that recent graduates are more instructionally effective, as measured by observation ratings, when their teachers serving as mentors modeled exemplary instruction and provided more frequent and higher quality coaching.

Preservice clinical evaluations are likely the first feedback that teacher candidates receive based on practice in authentic settings, but nearly all of the literature surrounding the impact of evaluations on teacher growth comes from the *inservice* context. Nonetheless, several studies provide evidence that structured feedback based on evaluations using formal rubrics can improve instructional practice (Burgess et al., 2021; Papay et al., 2020; Steinberg &

Sartain, 2015; Taylor & Tyler, 2012). For instance, Taylor and Tyler (2012) assess the roll out of a low-stakes evaluation policy and find that it increased participating teachers' effects on student achievement by about 0.10 standard deviations. Two recent randomized control trials of peer evaluation and coaching interventions have found similar results (Burgess et al., 2021; Papay et al., 2020).

One important concern about the viability of coaching interventions in the context of clinical placements is the quality and consistency of feedback provided by evaluators (Borko & Mayfield, 1995; Clarke et al., 2014; Matsko et al., 2018). Although formal preservice evaluations by mentors and field instructors do appear to predict workforce outcomes (Bartanen & Kwok, 2020; Chen et al., 2019), Bartanen and Kwok (2020) find that they exhibit poor interrater reliability. Indeed, a critique of university-based teacher education is that there is often little oversight about who is selected to serve as a mentor and whether mentor teachers are provided sufficient training to evaluate and provide feedback on clinical placements (National Council on Teacher Quality, 2017). Whether the findings on inservice feedback interventions generalize to preservice clinical teaching experiences is therefore an open question.

A growing body of research suggests that improving the quality of teachers serving as mentors and placements—measured by value added, experience, or school climate—can improve the quality of coaching received by teacher candidates. Ronfeldt, Brockman, and Campbell (2018; Ronfeldt, Matsko, et al., 2018) find positive correlations between the observational ratings of mentor teachers and the teacher candidates they mentor, who go on to become teachers. Similarly, both Ronfeldt, Brockman, and Campbell (2018) and Goldhaber, Krieg, and Theobald (2020) find that the effectiveness of mentor teachers (measured by value added) is associated with the instructional effectiveness of their mentees who enter teaching. And, in an experimental evaluation of the effects of assigning teacher candidates to more effective mentor teachers, Ronfeldt, Bardelli, et al. (2020) find that candidates placed with more instructionally effective mentors reported receiving higher quality and more frequent coaching and feedback.

Although largely nonexperimental, the emerging literature on preservice clinical experience offers evidence that placement and mentor teacher quality is important for teacher candidates and that preservice clinical evaluations provide teacher candidates a signal about their potential instructional effectiveness. In this study, we test the relationship between preservice clinical experiences and the within candidate growth of clinical evaluations using multiple evaluations of teacher candidates during a single clinical placement in a randomized controlled trial of placement procedures in two TEPs.

Background on the Improving Student Teaching Initiative

The Improving Student Teaching Initiative (ISTI), the experiment from which we derive the data used in this study, was designed to test the effects of different levels of feedback and the quality of clinical placements on teacher candidates' perceptions of their preparedness to teach, preservice clinical evaluations, work plans, and later workforce outcomes. Here we leverage data from one of two experiments, the "Placement Initiative," in which teacher candidates in two TEPs in Florida and Tennessee were randomly assigned to higher or lower quality clinical placements, as judged based on an index of mentor teacher and placement school characteristics (described in greater detail below).[2]

The implementation of the Placement Initiative included two key features: within TEP randomization[3] and the collection of clinical practice evaluations of teacher candidates by mentor teachers and field instructors. The Placement Initiative was designed to isolate a causally rigorous effect of placement quality on the growth in a teacher candidate's skills. Estimating the relationship between clinical placement characteristics and teacher candidate outcomes without random assignment could be biased by the matching of candidates and placement sites. Krieg et al. (2016), for instance, find evidence that teacher candidates who have higher basic skills licensure test scores, which are taken before clinical placements, are more likely to be matched to mentor teachers who also have higher basic skills licensure test scores.

To account for the concern about the matching process, we randomized teacher candidates to one of two placement lists. We first asked programs to identify the number of teacher candidates who needed placements in each district, grade, and subject combination. We refer to these as "placement blocks" since these are the blocks in which randomization eventually occurred. After the number of teacher candidates in each placement block was identified, TEPs asked districts to overrecruit potential mentor teachers to both ensure that there would be enough teachers for randomization to be successful and to offset any mentor teacher attrition. We then matched the host schools and potential mentor teachers to state administrative databases, using data prior to the 2016–2017 school year (the first year of the experiment), and ranked placements based on three mentor teacher attributes (experience, observational ratings, and a state-supplied measure of teacher value added) and two attributes of schools (school value added and retention rates).[4]

The above attributes of mentor teachers and schools were chosen based primarily on a growing literature that suggests associations between teacher candidates having better placements according to these attributes and measures of their later performance as early career teachers. For instance, teacher candidates with mentor teachers that had higher observational ratings also received higher observational ratings in their early career (Ronfeldt,

Brockman, & Campbell, 2018; Ronfeldt, Matsko, et al., 2018). Early career teachers received higher value-added ratings when the teachers who served as their mentors during the clinical portion of their education also had high value-added ratings (Goldhaber, Krieg, & Theobald, 2020; Ronfeldt, Brockman, & Campbell, 2018). And early career teachers were more likely to both have higher value-added estimates and remain in the workforce if they did their clinical practice in schools with higher average rates of teacher retention (Goldhaber, Krieg, & Theobald, 2017; Ronfeldt, 2012). Finally, recent graduates are more instructionally effective when they completed their clinical placements in schools with higher school-level value added (Bastian et al., 2020; Ronfeldt, 2015).

We divided clinical placements within each block into "higher" and "lower" quality lists based on whether an overall placement index, constructed from the above teacher and school attributes and described in the Data and Sample Description section, was above or below the median value index for a given block. In this article, we refer to the higher list as being higher quality and having more "instructionally effective" mentor teachers and higher quality field placement schools. But, to be clear, we do not have any direct measures of the quality of placements; when we refer to "higher" or "lower" quality placements we are referring to being above or below the median value of the placement index. Henceforth we refer to "higher quality" or "lower quality" placements as higher-index or lower-index placements, respectively. That said, the characteristics of mentors and field placement schools used to construct the index have all been shown to positively predict graduates' instructional effectiveness.

After identifying the two potential placement lists for each block, we randomly assigned teacher candidates to either a list of higher- or lower-index placements and returned the two sets of lists to the TEPs. The TEPs then assigned candidates to a specific placement on the assigned list, but we did not request any restrictions on how they made those assignments. As we discuss below, there are significant differences in the school and mentor teacher characteristics between these two lists and evidence of significant differences in treatment.

Data and Sample Description

Data Sources and Sample

The Placement Initiative within ISTI was implemented over two school years, 2016–2017 and 2017–2018, in two TEPs (in Florida and Tennessee); in total it included 398 teacher candidates (95% of whom attended the Tennessee TEP). Of these teacher candidates, 14 candidates randomized to the higher-index list and 15 randomized to the lower-index list were delayed in the program, did not meet residency requirements, changed majors, or

could not be linked to clinical evaluation data, resulting in an analytic sample of 369 candidates. These candidates were dropped from the experiment before beginning their clinical placements, and their attrition is not likely to be caused by exposure to higher- or lower-index placements.

We obtained data on mentor teachers from TEPs and state administrative data systems. TEPs provided a list of potential mentor teacher recruits with names and school-provided email addresses. Using this identifying information, the Florida and Tennessee Departments of Education matched these potential mentor teachers to state administrative and evaluation data sets. These data sets include information on teacher-level measures of value added, observational ratings, and experience, as well as school-level measures of value added and teacher retention. We have observational ratings of teachers serving as mentors as part of their inservice evaluations, as well as ratings by both field instructors and mentor teachers of teacher candidates. To distinguish between these, we henceforth refer to observational ratings of mentor teachers as "observational ratings" and of teacher candidates as "clinical evaluation ratings."

In Tennessee, the Tennessee Value-Added Assessment System (TVAAS) is used as the value-added measure, and in Florida, the state uses the Florida Value-Added Model (FVAM).[5] Teacher observational ratings in Tennessee are conducted multiple times throughout the year, and the evaluative rubric is provided by the Tennessee Educator Acceleration Model (TEAM).[6] Observational ratings in Florida are developed by districts in accordance with Florida's "Educator Accomplished Practices" (Florida Administrative Code, 2011), and we utilize the mentor teachers' final district observational rating on a 4-point scale. These same administrative records were used to construct average school value added and school level measures of teacher retention.[7]

As described in the Background on the Improving Student Teaching Initiative section, we employ these data to construct a placement index and leverage this index to build two lists—one for higher and one for lower quality clinical placements—across which teacher candidates were randomized. We combined these attributes into teacher and school indices. We standardized the individual mentor teacher attributes (teacher value added, experience, and observational ratings) to have means of zero and standard deviations of one and then summed them to create the teacher index. As these attributes are often cited as measuring different dimensions of a teacher's "ability," it is unclear whether certain attributes should be weighted more heavily. Although TEPs, states, and districts may value the dimensions differently, we choose to weight these teacher characteristics and measures equally when creating the teacher index. Similarly, we created a school index by summing standardized school average value added and teacher retention measures. In each case, the value-added measures are the official measures calculated by the state for teacher evaluation or school accountability purposes. We combined the teacher and school indices into an overall clinical

placement quality index, applying a weight of 0.75 to the teacher index and 0.25 to the school-level index. The teacher index was weighted more heavily because teacher candidates spend most of their clinical placements in their placement classrooms under the direct supervision of their mentor teachers. Hence, we hypothesized that mentor teachers would likely have a more proximal and potent impact on teacher candidates than their placement schools, and this hypothesis is consistent with observational research on the relative importance of mentor and school-level predictors of teacher candidates' early career effectiveness (Goldhaber, Krieg, Naito, & Theobald, 2020). It is this index that we used for within block randomization.

The evaluation data used as focal outcomes for this study come from the clinical evaluations of teacher candidates by field instructors and were collected from TEPs participating in the Placement Initiative. The programs collected this information through web-based teacher evaluation tools and provided spreadsheets containing interim and final sets of clinical evaluation ratings each term. We linked these data to random assignment information we had compiled before each placement began. Teacher candidates were observed by field instructors up to three times during the clinical placement. The analytical data set consists of the full panel of evaluations and contains multiple observations for each teacher candidate.

There is currently limited evidence about the reliability or external validity of preservice clinical evaluations. One important limitation is that evaluations, both inservice and preservice, appear to partially reflect classroom assignments or other features unrelated to the teacher's ability (Campbell & Ronfeldt, 2018, Cowan, Goldhaber, et al., 2020; Steinberg & Garret, 2016; Whitehurst et al., 2014). There is also some evidence that evaluations reflect the biases of evaluators against men or teachers of color (Bartanen & Kwok, 2020; Campbell, 2020; Steinberg & Sartain, 2020). In our application, rater prejudice is less concerning as a threat to validity given the randomization of teacher candidates to placement lists. Balance in teacher candidate demographics across higher- and lower-index placements would help ensure that any evaluator biases are similarly balanced. Two recent studies suggest that clinical evaluations exhibit significant rater error. Bartanen and Kwok (2020) find that only about 20% of the variance in clinical evaluation ratings is attributable to differences in teacher candidate quality and roughly 40% of the variance reflects disagreement among evaluators. Similarly, Gitomer et al. (2021) find significant rater error on the Educative Teacher Performance Assessment (edTPA). If placement quality does not influence rater biases, then the relatively high rater error in these measures will tend to diminish our treatment effects relative to the effect of the placement on true, unmeasured teacher quality. And, as we describe below, we show that the our results are robust to potential confounding effects of placements or instructor biases on clinical evaluation ratings.

Despite these potential shortcomings, there is evidence that clinical evaluations of instructional effectiveness are predictive of future effectiveness in the classroom. Two recent studies examine the predictive validity of the edTPA (a widely utilized clinical evaluation assessment completed during a teacher candidate's clinical placement). Bastian et al. (2018) find a one standard deviation increase in edTPA scores are associated with an 18% increase in teachers' first year value-added scores. Goldhaber, Cowan, and Theobald's (2017) findings are more mixed: performance on the edTPA is predictive of a teacher's future value added in math but is not statistically significant in reading.[8] Chen et al. (2019) investigate the predictive validity of another preservice clinical evaluation system, the Massachusetts Candidate Assessment of Performance (CAP), a state-wide clinical evaluation rubric partially aligned to the state's inservice teacher evaluation framework. They find that performance on this assessment predicts inservice observational ratings. Importantly, the clinical evaluation rubrics we use in this study are similarly aligned with inservice observational rating rubrics. And Bartanen and Kwok (2020) find that clinical evaluations are associated with other proxies for teacher candidate quality (e.g., high school GPAs, certification exam scores, pedagogical exam certification scores, and employment as a teacher). These findings are also generally consistent with the literature on observational ratings of teacher quality (Campbell, 2014; Harris & Sass, 2014; Kane et al., 2011; Kraft et al., 2020; Whitehurst et al., 2014).

Table 1 reports the number of teacher candidates participating in the Placement Initiative, as well as summary statistics by higher- or lower-index placements (Panel A). The differences between the teacher candidate characteristics in the higher- and lower-index placements are generally quite small, suggesting that the randomization was successful. We formally test this using an omnibus balance test between the higher- and lower-index preexperiment teacher candidate characteristics present in Panel A.[9] The p value of .262 reveals no significant differences between teacher candidates across higher- or lower-index placements.

Teacher candidates were randomized to higher- or lower-index placements, but our findings could still be confounded if there were systematic sorting of *field instructors* to the higher- and lower-index placements. For instance, if field instructors who are "tough graders" when it comes to clinical evaluation ratings are matched to mentor teachers who have more teaching experience, we would see a downwardly biased relationship between mentor experience (a component of the index) and field instructor ratings. Conversely, if field instructors who are "high growth" oriented are disproportionately matched to higher-index placement teacher candidates, this would positively bias the estimate of the experimental effect. We have no reason to believe this occurred, but to more formally address this concern, in Panel D of Table 1 we conduct an additional omnibus balance test of field instructors using field instructor indicators by the higher- or lower-index status of their

Table 1
Summary Statistics and Balance Tests

		Placement Status			
	Overall	Lower Index	Higher Index	Difference	Effect Size
Panel A: Teacher candidate characteristics					
Female	0.856	0.833	0.878	0.046	0.129
Male	0.141	0.16	0.123	−0.037	0.107
White	0.961	0.966	0.957	−0.008	0.044
Non-White	0.039	0.034	0.041	0.006	0.033
Cumulative GPA	3.61	3.625	3.598	−0.027	0.085
Inclusive GPA	3.435	3.411	3.457	0.046	0.129
Panel B: Mentor teacher and school characteristics					
Mentor value added	0.017	−0.399	0.361	0.760	0.890***
Mentor experience	0.009	−0.516	0.503	1.018	1.143***
Mentor obserbation rating	0.374	0.006	0.719	0.713	1.027***
School value added	0.005	−0.137	0.137	0.274	0.270**
School stay ratio	0.039	−0.023	0.107	0.130	0.131
Student teachers		180	189		
Panel C: Student characteristics at the teacher candidate school level					
Percentage					
African American	0.059	0.065	0.054	−0.010	0.155*
Asian	0.017	0.017	0.018	0.000	0.010
Hawaiian or Pacific Islander	0.002	0.001	0.002	0.000	0.020
Hispanic	0.106	0.113	0.100	−0.013	0.109
Native American	0.003	0.003	0.003	0.000	0.065
White	0.809	0.797	0.820	0.023	0.147
Eligibible for FRPL	0.373	0.395	0.353	−0.043	0.298***
	Chi-Square	*df*	*p*		
Panel D: Balance tests					
Teacher candidate characteristics from Panel A					
Stratification by randomization block	6.61	5	.251		
Field instuctor indicators					
Stratification by randomization block	44.76	44	.44		

Note. With the exception of observational ratings in Tennesse which are standardized at the state level due to data availability, all mentor teacher characteristics were standardized within the State-specific sample. Deviations from a mean of 0 and a standard deviation of 1 are due to sample attrition or a lack of evaluation data by either mentor teachers or field instructors. Differences between mentor characteristics are tested using *T*-tests. FRPL = free or reduced-price lunch.
*$p < .10$. **$p < .05$. ***$p < .01$.

teacher candidates. Results from this test (p = .49) indicate that this type of sorting did not occur.[10] Moreover, the majority of field instructors (87%) were working with teacher candidates on both of the placement lists, so if there were a "growth" predisposition among some field instructors, it would likely apply to teacher candidates on both the higher- and lower-index lists.

Panel B reports the average mentor teacher characteristics by higher- or lower-index placements, and as designed, we see significant differences between the characteristics of teachers serving as mentors for teacher candidates on the lower-index placement list compared with the higher-index placement list. For instance, teacher candidates assigned to the higher-index placement had mentor teachers with average value-added measures, observational ratings, and experience that exceeded the average values of the lower-index placement mentor teachers by 0.7 standard deviations or more for each measure.[11] The differences in school characteristics between the higher- and lower-index placements are in the expected directions but are far smaller in magnitude than those of the mentor teacher characteristics; this is to be expected since, as noted above, the overall placement index was more heavily weighted by mentor teacher characteristics (3/4) than school characteristics (1/4).

In Panel C of Table 1, we report average school-level student characteristics by higher- or lower-index placements. Teacher candidates cannot be linked directly to students, but rather schools and mentor teachers, thus results in Panel C are reported at the teacher candidate-school cell level. Since schools composed one fourth of the index and mentor teachers three fourths, it is possible to have one teacher candidate on the lower-index list and another on the higher-index list linked to the same school. Higher-index placements tend to be in schools that have more higher-income students and fewer African American students, but not by large margins. For instance, 35.4% and 39.5% of students linked to higher- and lower-index placement lists, respectively, were eligible for free or reduced-priced lunch.

Differences in the overall placement index are presented in Figure 1, which shows a kernel density plot of the placement index by whether or not the placement was considered a higher-index placement for its geographic-grade-subject randomization block. The mean difference between the higher- and lower-index placements is about 0.6 standard deviations on the index value.

Finally, while not reported in the Table 1, it is worth noting that some of the mentor teacher characteristics are significantly correlated within mentor teachers. In particular, the correlation between standardized observation ratings and experience of mentor teachers is 0.25, and the correlation between mentor teacher observation ratings and value added is 0.28. The correlation between the value added and experience of mentor teachers was close to zero (−0.05) and statistically insignificant; this is not terribly surprising given the small sample of mentor teachers and the fact that mentor teachers were

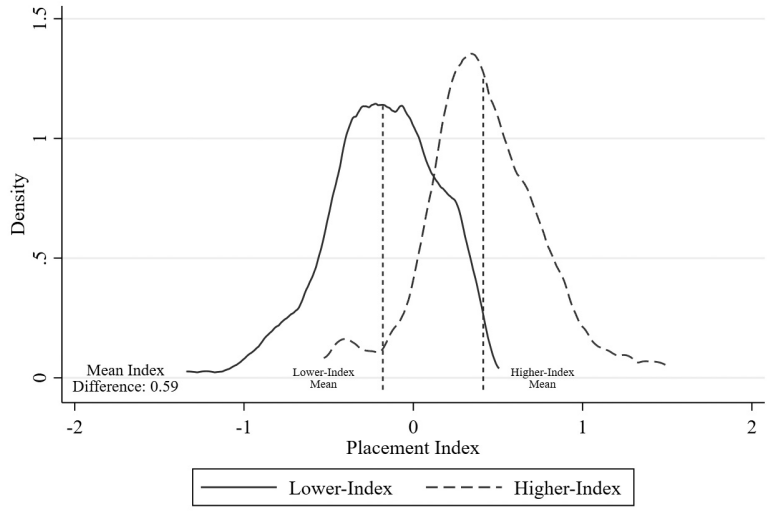

Figure 1. **Distribution of index values by placement status.**

Note. Figure 1 is a kernel density plot of the Placement Index for teacher candidates described in the Background on the Improving Student Teaching Initiative section. High-index teacher candidates were randomized to high-quality mentor teachers, that is, were treated. Randomization was done within geographic-grade-subject cells.

typically well into their careers and thus beyond the period where there is a rapid increase in effectiveness associated with gaining additional experience (Papay & Kraft, 2015; Rice, 2010).[12]

Clinical Evaluation Ratings

Clinical evaluations consist of ratings on a number of individual teaching skills (*items*) grouped into different instructional *domains*. The TEP rubrics from both programs have three instructional domains: instruction, environment, and either planning or assessment depending on the TEP. The two participating TEPs have 18 and 19 individual items nested within these domains (e.g., "Lesson structure and pacing," "Motivating students," and "Teacher content knowledge"). Field instructors in each TEP scored the items on a different 5-point Likert-type scale.[13]

The scales describe candidate performance either relative to a standard (e.g., above, below, at expectation) or in developmental terms (e.g., beginning, applying, and innovating). A teacher candidate receiving the highest evaluation rating on an item is considered "Significantly Above Expectations" in the Tennessee TEP and "Innovating" in the Florida TEP. The second to highest

rating uses the terms "Above Expectations" in Tennessee and "Applying [teaching skill] to classroom practice" (in Florida), whereas the bottom two categories of the scale used the terms "Significantly Below Expectations" and "Below Expectations" in Tennessee and "[teaching skill] Not Used" and "Beginning" (in Florida).

Both states are working to align expectations of teacher candidates with those of inservice teachers. As such, the clinical evaluation rubric for teacher candidates is aligned to the observational rubric used for inservice teachers. In Tennessee the TEP clinical evaluation rubric used is identical to the state's inservice observational rating rubric, the TEAM statewide observational rubric. In Florida, the clinical evaluation rubric is similar to Florida's Educator Accomplished Practices, "Florida's core standards for effective [inservice] educators" (Florida Department of Education, 2019). The TEP rubric shares the same item level scale as the sample evaluation rubric of inservice teachers provided by the Florida Department of Education.[14]

It is teacher candidates' clinical evaluation ratings by field instructors that are the focal outcomes of this study. In particular, in the analyses we describe below, we focus on whether placement quality (as measured by the placement index) and/or mentor or field placement characteristics affect the clinical evaluation ratings collected by field instructors. We have information on these clinical evaluations at the item level, which we standardize within TEP and cohort to have a mean of zero and a standard deviation of one. In total we have records for 369 teacher candidates (180 in lower-index placements and 189 in higher-index placements) who received an average of 2.9 clinical evaluations from field instructors throughout the course of their clinical placement. During each clinical evaluation, teacher candidates were rated on multiple items resulting in a total of 18,080 teacher candidate–clinical evaluation–item level records. Of this sample, 95% of the observations come from the much larger TEP in Tennessee.[15] Each of the roughly three evaluations per teacher candidate took place throughout the clinical placement, with the first evaluation taking place 6 weeks into the field placement and subsequent evaluations generally 16 to 20 weeks later; on average, the last evaluation took place 26 weeks into the field placement, or toward the end of a yearlong placement.[16]

In Figure 2, we depict the progression of the distribution of item-level evaluation ratings by field instructors over observation number (in other words, order of evaluation). The average ratings are relatively high in the sense that few use much of the range of the scale. In fact, we have combined the lowest two rating categories ("Significantly Below" and "Below Expectations") into one category for this figure on account of the lowest rating category comprising a very small proportion of clinical ratings.[17,18] Ratings are high from the first time that teacher candidates are observed. For instance, the first time that teacher candidates receive clinical evaluations, over 46% of teacher candidate items are rated "Above" or "Significantly Above" expectations. By the third and final

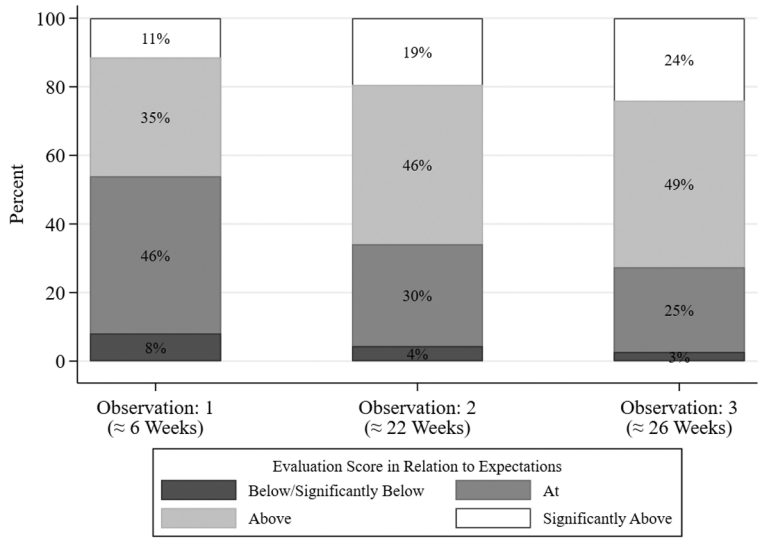

Figure 2. **Changes in evaluation ratings over the clinical placement.**

Note. Distribution of ratings by order of observation. We collapsed the bottom two categories into a single rating as so few teacher candidates received the lowest score (0.5% on the first observation). For one program that used a 5-point scale, but used software permitting ratings on a decimal scale, we took the largest integer smaller than the recorded score for any noninteger scores. Evaluation rubric category names are taken from the Tennessee teacher education program.

observation, more than 70% of teacher candidate-items are rated "Above" or "Significantly Above" expectations, while less than 3% are "Below" or "Significantly Below" expectations.[19]

Given the intended alignment between the clinical and inservice evaluation rubrics, this could imply that the great majority of teacher candidates are receiving the message that they are exceeding expectations on the great majority of teaching skills assessed by these rubrics before they are actually hired as teachers. On the other hand, field supervisors and mentor teachers may alter their standards for teacher candidates; for example, their criteria for "meets expectations" may be more lenient for teacher candidates who are just learning to teach.

Notwithstanding the generally high clinical evaluation ratings, we find field instructors are more likely to be "tougher graders" of teacher candidate than mentor teachers. In Supplemental Appendix A (available in the online version of the journal), we leverage our data to assess the ratings that teacher

candidates receive from both mentors and field instructors. In a model regressing evaluation ratings on an indicator for the type of rater (mentor or field instructor),[20] we find, after controlling for the timing of observation, mentor teachers are more likely to give the same teacher candidates significantly higher item evaluation ratings, by about 0.12 standard deviations. This result comports with the findings of Gareis and Grant (2014) that field instructors issue lower early clinical placement evaluations ratings to teacher candidates, relative to mentor teachers.

Analytic Approach

Does a higher quality clinical placement improve the teaching practice of candidates? We answer this question by leveraging the random assignment of teacher candidates to either higher- or lower-index placement lists, outlined in the Background on the Improving Student Teaching Initiative section. We implement an intent-to-treat analysis of the effect of placement list quality on evaluation ratings[21] with an indicator for assigned placement list, as in Equation (1):

$$y_{iltjb} = \beta_1 High_i + \sum_{s=1}^{3} \beta_{2,t} 1[Order_{itjb} = s] + \sum_{s=1}^{3} \beta_{3,t} High_i * 1[Order_{itjb} = s] + \mu_b + [\theta_j + \omega_i + \alpha_l + \varepsilon_{iltjb}] \quad (1)$$

In Equation (1), i denotes the individual teacher candidate, l denotes the item, t denotes the timing of the observation (the ordinal position of the rating), j denotes the field instructor, and b denotes the randomization block. $High_i$ is a binary indicator for whether or not a teacher candidate was randomized to the higher-index placement list. $1[Order_{itjb} = s]$ is an indicator for the observation from the sth ordinal position (by date) within candidate-item cells.[22] Teacher candidates were evaluated up to three times throughout their clinical placements, and here t indexes the first, second, or third evaluation. μ_b is a vector of placement block fixed effects. Equation (1) includes nested field instructor (θ_j), teacher candidate (ω_i), and item (α_l) random effects.[23,24] We adjust for individual and rater effects to account for correlated sources of error and for the fact that raters differ significantly in their assessment of individual candidates (Bartanen & Kwok, 2020).

We hypothesize that exposure to higher-index clinical placements may have a cumulative effect over time on teacher effectiveness due to a faster growth rate in skill acquisition. For instance, Gareis and Grant (2014) found that teacher candidates of more highly trained mentor teachers had faster growth in their clinical evaluation ratings over the course of their clinical placement. Outside of a failure of randomization or an unforeseen source of bias, the only way we should see differences in clinical evaluation ratings is if there are differences in growth. We therefore expect that effects of the

placement assignment on evaluations will be larger for evaluations that occur later in the term. We investigate this possibility by including in some specifications of Equation (1) an interaction term between $High_i$ and each of the $1[Order_{itjb}=s]$ indicators.

Randomization of candidates to placement lists ensures that candidate effectiveness is not correlated with higher- or lower-index placements, which would bias estimates of the coefficients in Equation (1). However, even with random assignment of candidates to placement lists, there are still three primary challenges to interpreting β_1 as the causally rigorous effect of higher-index placements on teaching practice.

First, by nature of the experiment, teacher candidates on the higher-index list are more likely to be assigned to higher achieving classrooms, and variation in classroom settings could contribute to the estimation of the higher-index placement coefficient in two ways. Higher achieving and/or better managed classrooms may be an easier environment for teacher candidates to learn new skills. On the other hand, field instructors could misread classrooms managed well by mentor teachers and attribute this to the teacher candidate. The first contribution would reflect an actual effect of higher-index placements while the second would introduce bias.[25] Prior research, which has found that teachers tend to receive higher observational ratings when they teach in higher achieving classrooms (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016), suggests that this bias is a possibility. To the extent that clinical placements from the higher-index result in more advantaged environments, candidates in such schools may systematically benefit in their clinical evaluations. In other words, this could potentially bias our estimate of β_1 upward.

To address this concern, we modify Equation (1) to include the field instructor's prior item-level clinical evaluation ratings, ($Y_{il(t-1)jb}$), which is depicted in Equation (2):

$$y_{iltjb}=\beta_1 High_i+\beta_2 Y_{il(t-1)jb}+\sum_{s=2}^{3}\beta_{3,t}1[Order_{itjb}=s]+\mu_b+\left[\theta_j+\omega_i+\alpha_l+\varepsilon_{iltjb}\right] \quad (2)$$

This specification removes the influences of the classroom environment on clinical evaluations reflected in the prior clinical evaluation (as well as any potential effects clinical placement quality had on teacher candidates through their prior clinical evaluation).[26] It therefore identifies the effect of higher-index clinical placements by comparing the growth rates of candidates assigned with higher- or lower-index clinical placements. In some specifications, we replace field instructor random effects with fixed effects.

Equation (2) ameliorates this potential upward bias so long as the influence of classroom context on measured (rather than actual) performance is fully accounted for in the prior field instructor evaluation. This would *not* be the case if field instructors place more weight on the clinical placement

quality as the clinical placement progresses or if clinical placement quality was perceived to improve over the course of the year (and was reflected in field instructor ratings of teacher candidates). We view these possibilities as unlikely because candidates tend to take on additional teaching responsibilities over the course of the clinical placement, which suggests that field instructors likely place more weight on the skills of the teacher candidates throughout the year. Nonetheless, we approach this potential source of bias by replacing the prior evaluation rating with a teacher candidate fixed effect. This controls for any time-invariant inherent ability of the teacher candidate as well as any potential source of bias from the classroom environment.[27] The overall effect of the higher-index placements on clinical evaluations is not identified in models with teacher candidate fixed effects. Thus, the coefficients of interest in these regressions are those on the interactions between order indicators and higher-index placement assignment, which provides an estimate of the additional improvement in clinical practice among candidates assigned to the higher-index placement list.

A second concern is that mentor teachers may influence the clinical evaluation ratings given by field instructors. Mentor teachers could, for instance, influence field instructor ratings by communicating their perceptions of teacher candidates to the field instructors. As the placement index used to construct higher- and lower-index placements was weighted to reflect the characteristics of mentor teachers, an estimated higher-index placement effect may partially reflect the evaluation standards of the mentor teachers. Although we cannot test this directly, in the Threats to Causal Interpretation section, we provide evidence that mentor teachers on the higher-index list do not provide ratings higher or lower than would be expected given the field instructor evaluations. In particular, we use a subset of candidates for whom we have a joint observation conducted by the field instructor and mentor teacher and regress the differences in ratings on the indicator for higher- or lower-index placement status and field instructor fixed effects. We do not find systematic differences in ratings across placement lists.

Last, although the assignment of mentor teachers was determined by the random assignment process, we did not control the assignment of field instructors. It is not clear how matching of field instructors to candidates would bias our results. Programs did not know which lists contained the higher-index placements, but they did understand the research design and it is possible they correctly identified the lists. While field instructors were not informed of which placements had higher- or lower-index values, and it is unclear the extent to which TEPs even informed field instructors that an experiment was ongoing, it is possible field instructors also deduced the different experimental conditions. If program administrators, for instance, compensated for less promising placements with field instructors who are better or more generous evaluators and/or field instructors modulated their feedback or clinical evaluation ratings to assist teacher candidates in lower-index

placements, our estimates would tend to be biased downward. Conversely, conditional on program administrators and field instructors discerning the different experimental conditions, a desire to see positive impacts for teacher candidates based on the higher-index experimental condition could bias our estimate upward. However, this does not appear likely in this case, especially since the vast majority of field instructors supervised teacher candidates on both lists. We show in Table 1 that individual field instructors are balanced higher- and lower-index groups. As an additional check, we replicate our main results from Equation (1) replacing the field instructor random effects with fixed effects. This specification compares teacher candidates assigned with the same field instructor, but across higher- and lower-index groups, which alleviates the concern of sorting on field instructor rating standards by higher- and lower-index status. Moreover, the models that include field instructors' own lagged clinical evaluations of teacher candidates should capture any potential field instructor bias toward the different experimental conditions.

As described above, the higher-index indicator is a composite measure built from both multiple mentor teacher characteristics and school placement characteristics. If it appears that higher-index clinical placements improve the teaching practice of teacher candidates, an important question is which specific dimensions of clinical placements improve practice. We attempt to disentangle the effects of higher-index clinical placements on teaching practice in the descriptive model shown below in Equation (3):

$$y_{iltjb}=\beta_{1,m}Index_{i,m}+\beta_2 High_i+\sum_{s=1}^{3}\beta_{3,t}1[Order_{itjb}=s]+\mu_b+\left[\theta_j+\omega_i+\alpha_l+\varepsilon_{iltjb}\right] \quad (3)$$

Equation (3) is similar to Equation (1) but now includes individual mentor teacher and school placement characteristics. These characteristics are included in the index ($Index_{i,m}$), where m, depending on the specification, denotes mentor teacher value added, experience, and observational ratings and school-level teacher retention and value added. We add variables from $Index_{i,m}$ individually to the regressions to see, what if any, effect each variable has on the estimated higher-index coefficient (β_2). If any of the mentor or placement attributes are missing, we create an indicator for missing values, impute the value as the Placement Block average,[28] and include both the indicator and imputed values in the index vector. As with Equation (1) in some specifications we interact evaluation order with the higher-index indicator ($High_i$).

Results

Leveraging the Experiment: Clinical Evaluations and Higher- and Lower-Index Placements

In Table 2, we report the intent-to-treat estimates using the random assignment of teacher candidates to clinical placement lists. Given the

Table 2

The Placement Initiative Experiment's Impact on Growth in Clinical Evaluation Ratings

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Second evaluation	0.372***	0.371***	0.373***	0.372***				0.368***
	(0.013)	(0.018)	(0.013)	(0.018)				(0.020)
Third evaluation	0.548***	0.520***	0.549***	0.521***		0.066***	0.067***	0.514***
	(0.013)	(0.018)	(0.013)	(0.018)		(0.014)	(0.014)	(0.020)
Prior clinical evaluation						0.298***	0.297***	
						(0.009)	(0.009)	
Higher-index placement	0.078*	0.061	0.082**	0.065	0.082*	0.063*	0.080**	
	(0.043)	(0.045)	(0.041)	(0.043)	(0.044)	(0.038)	(0.036)	
Higher-index placement × Second evaluation		0.003		0.002				0.005
		(0.025)		(0.025)				(0.028)
Higher-index placement × Third evaluation		0.057**		0.057**				0.059**
		(0.026)		(0.026)				(0.029)
Field instructor fixed effects			×	×			×	
Field instructor-observation fixed effects					×			
Teacher candidate fixed effects								×
Placement block fixed effects	×	×	×	×		×	×	
Observations	18,080	18,080	18,080	18,080	18,080	11,526	11,526	18,080
Teacher candidate–mentor pairs	369	369	369	369	369	340	340	369
Field instructor effects variance	0.133	0.133	0.237	0.238		0.080	0.227	0.000

Note. Regressions of standardized clinical evaluation ratings on order of evaluations and placement status. Order indicates the observational order of the evaluation within candidate-item-field instructor cells. All evaluations (the dependent variable) were submitted by field instructors. All models include field instructor, candidate, and item by program random effects, and an indicator for participation in the feedback inititive.
*$p < .10$. **$p < .05$. ***$p < .01$.

standardization of the clinical evaluations, the coefficient estimates should be interpreted as the estimated standard deviation change in field instructor clinical evaluation ratings of teacher candidates on the higher-index placement list relative to the evaluations of teacher candidates on the lower-index placement list.[29]

We begin with a specification (column 1) that just includes a higher-index indicator (equal to 1 if a candidate is in the higher-index group). The indicator is statistically significant at the 10% level, and the estimate suggests that teacher candidates in the higher-index group tend to receive higher field instructor clinical evaluation ratings throughout their placements.

In the next column, we consider candidate growth over the placement. Specifically, we add interactions between evaluation order and the higher-index indicator. We find that the third and final evaluation that teacher candidates receive is significantly higher, by about 6% of a standard deviation, for those in the higher-index relative to the lower-index groups. Given that the average change between the first and final evaluations for all teacher candidates was 52% of a standard deviation, a 6% increase of a standard deviation represents approximately 12% of the average total change in clinical evaluation ratings. In this interaction model, the cumulative effect of the higher-index placements should be interpreted as the main effect, "higher-index placement," plus the interaction, "higher-index placement × third evaluation," which results in a treatment effect by the end of the clinical placement of roughly 12% of a standard deviation. Consistent with the findings reported in Figure 1, there is clear evidence of increased ratings from the first to third evaluation of teacher candidates.[30] We more accurately depict this growth in Figure 3. Figure 3 is a linear combination of the coefficients of the constant, evaluation order indicators, higher-index indicator, and interactions between the higher-index indicator and evaluation order indicators from the model estimated in column (2) of Table (2) with 90% confidence intervals. Overlapping confidence intervals appear to be driven by the variance on the constant and order indicators, which do not enter into the difference.

In columns (3) and (4) we replace field instructor random effects with field instructor fixed effects. The random effects specification assumes that field instructor rating patterns are uncorrelated with higher- or lower-index assignment. Since field instructors tend to have candidates on both higher- and lower-index lists, we believe this assumption is plausible. Nonetheless, if some field instructors assume responsibility for teacher candidates in particular schools or regions, the assignment procedures might introduce bias through differential rater effects. In column (5), we replace these with field instructor-by-observation order fixed effects. Including the observation order in the grouping allows for the possibility that higher-index candidates have more "growth-oriented" field instructors whose clinical evaluation ratings improve more quickly over time. Estimates are quite similar to the baseline estimates, suggesting that there is not much sorting of field instructors to

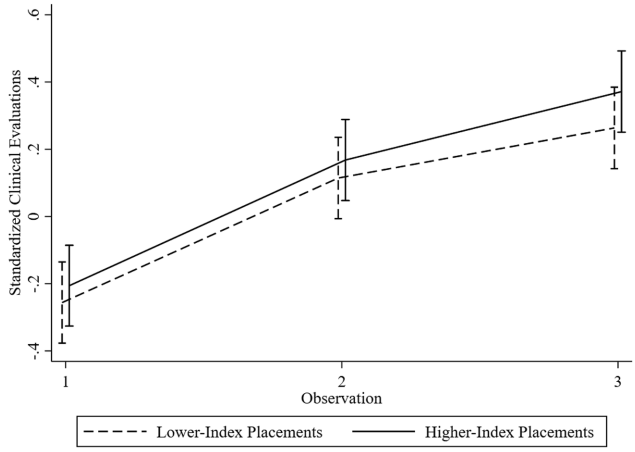

Figure 3. **Growth in clinical evaluation ratings by placement list.**
Note. This figure is a linear combination of the coefficients of the constant, order, higher-index placement indicator, and interactions between the placement indicator and order variables from the model estimated in column (2) of Table (2) with 90% confidence intervals.

higher- or lower-index list candidates in our experiment, consistent with reports from programs about how they assigned field instructors.

In columns (6) and (7), we display results for models that use outcome data from the second and third observations and control for the prior field instructor evaluation rating. Given the adjustment for prior clinical evaluation ratings in these models, the coefficient on the higher-index placement should be interpreted as an effect on teacher candidate growth during the clinical placement. If the apparent effects of higher-index placements on clinical evaluations were caused by a favorable classroom environment created by the mentor teacher, we would expect that adjusting for a prior clinical evaluation of the candidate *in the same classroom* would attenuate the coefficient on higher-index placements. The coefficients on the higher-index placement are 0.06 to 0.08 standard deviations. These estimates are consistent with those shown in columns (1) through (5). In column (8), we include one more robustness check that controls for a teacher candidate fixed effect, and these results are consistent with columns (6) and (7). This suggests that candidates on the higher-index list improve their measured teaching skills more than candidates on the lower-index list throughout the clinical placement.

Both the random and fixed effects specifications of the model permit us to recover the specific field instructor effects. One can interpret these fixed effects as a measure of how harsh or lenient an evaluator a field instructor

is, on average, when evaluating candidate growth. This is a bit speculative in that the field instructor ratings would presumably reflect both the ratings standards of field instructors as well as their effects on the instructional skills of teacher candidates. That said, we estimate the variance of the field instructor random effect to be 0.133. Because the ratings are standardized, this means that differences in standards, differences in candidate quality, and/or impact of coaching across field instructors account for about 13.3% of the variation in candidate ratings.[31] Although we cannot distinguish the effects of field instructors on candidate practice from differences in their evaluation standards, these estimates are comparable to the variance of rater effects on observational rubrics found in the Measures of Effective Teaching Project (Ho & Kane, 2013).

Exploring the Mechanisms of the Placement Effect

The evidence presented in the prior section suggests that the quality of the placement matters for teacher candidate growth in clinical practice. But what specifically about the placement affects teacher candidates? In this section, we present an exploratory analysis (Equation 3) of the impact of different placement characteristics (the experience, value added, and observational ratings of teachers serving as mentors, as well as school-level retention and value added) on the higher-index coefficient.

In Table 3, we add each of the characteristics of mentor teachers (columns 3–8) and school characteristics (columns 9–12) to the model.[32] We interpret the diminishment of the higher-index indicator with the inclusion of a mentor or school characteristic as indicative of the importance of the particular characteristic given that the model is then showing the influence of the higher-index placement after controlling for the characteristic in question. For example, if the inclusion of mentor teaching experience results in a substantial decrease in the magnitude of the estimated higher-index effect and a positive coefficient on experience, we would interpret that as mentor teaching experience being a vital component of the higher-index construction (i.e., treatment effect). If on the other hand, the higher-index effect is unchanged, then we would interpret that as mentor teacher experience contributing little, if anything, to an estimated higher-index effect.[33]

For comparison, we include columns (1) and (2) from our main models in Table 2. The inclusion of the mentor teacher value added (columns 3 and 4) leaves the estimated higher-index effect nearly unchanged. The estimate on value added is negative, but not statistically significant. We note that value added is missing for many mentor teachers, rerunning the model with only complete data (i.e., listwise deletion) produces a positive but insignificant coefficient estimate for value added. The inclusion of experience in the model as a covariate significantly diminishes the magnitude of the coefficient on higher-index placements in column 5 (from 0.078 to 0.022) and, in particular, the clinical evaluation ratings that teacher candidates receive on the first

Table 3

Placement Initiative Experiment, Placement Characteristics, and Clinical Evaluation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Second evaluation	0.372***	0.371***	0.372***	0.371***	0.372***	0.371***	0.372***	0.371***	0.372***	0.371***	0.372***	0.371***
	(0.013)	(0.018)	(0.013)	(0.018)	(0.013)	(0.018)	(0.013)	(0.018)	(0.013)	(0.018)	(0.013)	(0.018)
Third evaluation	0.548***	0.520***	0.548***	0.520***	0.548***	0.520***	0.548***	0.520***	0.548***	0.520***	0.548***	0.520***
	(0.013)	(0.018)	(0.013)	(0.018)	(0.013)	(0.018)	(0.013)	(0.018)	(0.013)	(0.018)	(0.013)	(0.018)
Higher-index placement	0.078*	0.061	0.076*	0.059	0.022	0.005	−0.000	−0.017	0.075*	0.058	0.076*	0.059
	(0.043)	(0.045)	(0.044)	(0.046)	(0.049)	(0.051)	(0.049)	(0.051)	(0.043)	(0.045)	(0.043)	(0.045)
Higher-index placement × second evaluation		0.003		0.003		0.003		0.003		0.003		0.003
		(0.025)		(0.025)		(0.025)		(0.025)		(0.025)		(0.025)
Higher-index placement × third evaluation		0.057**		0.058**		0.057**		0.057**		0.057**		0.057**
		(0.026)		(0.026)		(0.026)		(0.026)		(0.026)		(0.026)
Mentor value-added			−0.013	−0.012								
			(0.036)	(0.036)								
Mentor experience (*SD*)					0.060**	0.060**						
					(0.027)	(0.027)						
Mentor standardized observational rating (*OR*)							0.115***	0.115***				
							(0.039)	(0.039)				
School value added									0.007	0.007		
									(0.031)	(0.031)		
School retention											0.016	0.016
											(0.038)	(0.038)
Placement block fixed effects	×	×	×	×	×	×	×	×	×	×	×	×
Observations	18,080	18,080	18,080	18,080	18,080	18,080	18,080	18,080	18,080	18,080	18,080	18,080
Teacher candidate–mentor pairs	369	369	369	369	369	369	369	369	369	369	369	369
Field instructor effects variance	0.133	0.133	0.135	0.135	0.139	0.139	0.130	0.131	0.133	0.134	0.131	0.132

Note. Regressions of standardized clinical evaluation ratings on order of evaluations and placement status. Order indicates the observational order of the evaluation within candidate-item-field instructor cells. All evaluations (the dependent variable) were submitted by field instructors. All models include field instructor, candidate, and item by program random effects, and an indicator for participation in the feedback inititive.
*$p < .10$. **$p < .05$. ***$p < .01$.

evaluation by field instructors (column 6). This is consistent with the positive and significant coefficient on mentor teaching experience. In columns 7 and 8, we include the effects of the inservice observational ratings of mentor teachers. The coefficient on the higher-index placement is significantly attenuated with the inclusion of observational ratings, which is consistent with the positive and significant estimates on the observational ratings (columns 7 and 8). When we include mentor observational ratings, the magnitude of the coefficient on higher-index placements drops to zero.

Finally, in columns 9 to 12, we show the inclusion of school characteristics in the model. The findings on the higher-index placement coefficient are largely unaffected by the inclusion of either of the two school-level characteristics (teacher retention and school level value added) that are components in the placement index. For instance, the magnitude of the higher-index placement coefficient is reduced by only 4% for the school value-added model and 3% for the school retention model. This suggests that these school characteristics explain little of the effect of the higher-index placements in influencing the clinical evaluation ratings. (Importantly, the school-level characteristics only comprised a quarter of the weight of the overall index; as such, we might not expect them to have a big effect on the higher-index indicator.)

By including the higher-index indicator in Table 3, which is consistent with Equation (3), we are implicitly comparing mentor teacher and school characteristics within higher- or lower-index lists. That is, we are only capturing the within placement index list category variation of mentor characteristics. In Supplemental Appendix B (available in the online version of the journal), we report results that omit the higher-index indicator so that the results are based on both within and across higher- and lower-index group variation in mentor teacher and school characteristics (we also estimate models with and without mean imputed placement characteristics). Results are quite similar.

In Supplemental Appendix C (available in the online version of the journal), we further develop a quasi-experimental approach for ascertaining the effects mentor teacher characteristics have on clinical evaluations by comparing the within placement block treatment/control differences in clinical evaluations and mentor teacher characteristics across blocks. That is, if the coefficients on mentor characteristics are causal, then randomization blocks with greater average differences in mentor characteristics across randomization lists should also have larger differences in average field instructor ratings. Results are directionally consistent for mentor value added and experience, but the results are mixed for observational rating depending on the specification. No coefficients are statistically significant, which is likely the result of low power.

Threats to Causal Interpretation

The relationship between higher-index placements and teacher candidate performance appears robust to two of the three threats to identification

raised in the Analytic Approach section. Candidates assigned higher-index placements earned higher clinical evaluation ratings overall and developed more quickly over time. We also find little evidence that sorting of mentors or teacher candidates to field instructors explains our results. The remaining concern is that differences in rating standards across *mentor teachers* might influence the ratings provided by field instructors. That is, higher quality mentor teachers may systematically rate candidates higher or lower conditional on unobserved ability and thereby influence field instructor ratings without affecting candidates' teaching skills. We investigate this possibility by comparing field instructor and mentor teacher ratings for the same teacher candidate on the same clinical evaluation.

We use data on joint observations by field instructors and mentor teachers to test this possibility. We difference observations of the same clinical evaluation rubric item and observation window to remove any time or candidate effects and focus only on differences between the mentor and field instructor rater effects. That is, we regress

$$y_{iltjbp} - y_{iltj^*bp} = \beta_m Index_{i,m} + \tau_j + \varepsilon_{iltjbp}$$

where y_{iltjbp} is teacher candidate i's clinical evaluation rating at TEP p in block b on item l during observation t given by field instructor j and y_{iltj^*bp} is the analogous clinical evaluation rating but given by their mentor teacher. τ_j is a field instructor fixed effect. Depending on the model, $Index_{i,m}$ is either the indicator for the higher-index placements or individual mentor teacher characteristics. A positive coefficient on $Index_{i,m}$ indicates that higher quality mentors provide systematically higher ratings. All error terms are clustered at the teacher candidate level.

We present the results of these tests in Table 4 for evaluations conducted within the same week (columns 1–4) and on the same day (columns 5–8). The coefficients on mentor characteristics are imprecisely estimated and not consistently signed. The only statistically significant finding is that higher value-added mentor teachers rate their candidates lower than field instructors. If that is the case, and if field instructors take mentor perceptions into account when formulating their own ratings, then our estimates would tend to be biased downward; that is, our experimental results would be a lower bound on the effect of higher-index placements. More experienced mentor teachers and those on the higher-index lists rate their candidates higher, although the differences are not significant. The sign of the coefficient on the observational ratings of mentor teachers depends on whether we use joint observations or those submitted the same week. Overall, we do not find consistent evidence of systematic variation in mentor evaluations associated with their observable teacher characteristics, although the confidence intervals cannot rule out substantial differences.

Table 4

Differences in Clinical Evaluation Ratings by Field Instructors and Mentor Teachers

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Experiment	Higher-index placement	0.058				0.004			
		(0.107)				(0.098)			
Mentor characteristics	Mentor value added		−0.130**				−0.065		
			(0.062)				(0.041)		
	Mentor experience (*SD*)			0.017				−0.008	
				(0.050)				(0.052)	
	Mentor standardized observational rating (*OR*)				−0.014				0.048
					(0.076)				(0.068)
	Placement block fixed effects	×	×	×	×	×	×	×	×
	Observations	2,023	2,023	2,023	2,023	1,726	1,726	1,726	1,726
	Teacher candidate–mentor pairs	76	76	76	76	66	66	66	66
	Clinical evaluations submitted within *x* days	7	7	7	7	0	0	0	0

Note. Regression of the difference between mentor teacher clinical evaluations and field instructors during joint observations on placement status and mentor characteristics. All models include field instructor fixed effects and are clustered at the teacher candidate level. Joint observations are identified by the number of days between field instructor and mentor teacher evaluation submissions.
*$p < .10$. **$p < .05$. ***$p < .01$.

Discussion and Conclusions

To our knowledge, this article is the first to document whether development of observed teaching skills while in clinical practice is causally attributable, at least in part, to the characteristics of mentor teachers and field placements schools. The results suggest that higher quality field placements (as measured by the higher-index indicator) cause candidates' instructional effectiveness to grow at a faster rate during clinical placements. This finding is consistent with prior correlational evidence but provides some of the only causally rigorous evidence for the importance of high-quality field placements during initial teacher preparation. While not completely analogous to inservice observational ratings, the 0.06 standard deviation increase between the first and third clinical evaluations caused by higher-index clinical placements is 73% of the gains seen in the inservice observational ratings of teachers due to an additional year of teaching experience (Kraft et al., 2020).

Prior work has found that teacher candidate placements can be ad hoc; TEPs work with district offices, principals, or even reach out to individual mentor teachers to place their students (St. John et al., 2018). Placement coordinators at TEPs report keeping track of mentor teachers they believed to be particularly effective (St. John et al., 2018). The ad hoc nature of the matching process plus our findings on the importance of higher quality placements suggests that improving the matching process for teacher candidates could lead to measurable skill increases. To that end, Ronfeldt, Bardelli, et al. (2020) finds that when TEPs are provided information on the instructional effectiveness of mentor teachers they act on it and select higher quality mentors for their teacher candidates. And there may be lots of scope to change the matching process, as only 3% of teachers serve as mentor teachers a year (Goldhaber, Krieg, & Theobald, 2020). In fact, the research design of this study encouraged programs to overrecruit for placements, which they did without much difficulty. As such, it is not impossible to find higher quality placements even under normal recruitment operations, provided that TEPs possess the knowledge of which placements are likely higher quality.

In light of the findings on growth in teaching skills during clinical practice, state policymakers may wish to take active roles in shaping clinical placements. While states typically establish requirements for the time that teacher candidates need to spend in clinical placements, few states have detailed requirements for who can serve as a mentor teacher. For instance, only 20% of states require mentor teachers to have 3 years of experience or more (Greenberg et al., 2011), and only 32% require that mentor teachers meet specific performance benchmarks (National Council on Teacher Quality, 2020).[34] Our findings suggest that such policies could improve the instructional effectiveness of early career teachers.

Policymakers may worry that restrictions on clinical placements could disadvantage schools serving low income or underrepresented minority

(URM) students. Higher income and White students tend to have more effective teachers (Goldhaber et al., 2019), so policies that select mentors based on their effectiveness or observable characteristics might reduce the diversity of clinical placements. Teacher candidates also frequently obtain positions in their placement schools (Krieg et al., 2016), and policies that limit placement in hard to staff schools could potentially exacerbate the challenges of hiring teachers.

However, a careful reading of the research on school and teacher effectiveness suggests states can simultaneously improve the effectiveness of mentor teachers and the diversity of clinical placement sites. As we mentioned above, there is a large, untapped pool of potential mentor teachers (3% of teachers serve as mentor teachers in any given year). Although there are gaps between URM and White students, and between economically disadvantaged students and noneconomically disadvantaged students, in their access to high quality teachers (Goldhaber et al., 2019), average teacher effectiveness varies relatively little across schools (Mansfield, 2015; Rivkin et al., 2005). Thus, preparation programs should be able to recruit effective mentor teachers in a wide variety of settings. Similarly, effective schools and schools with strong professional climates can be found in all kinds of neighborhoods and districts. In Tennessee in the 2018–2019 school year, 39% of schools with a high level of URM students (above the median) received either of the top two marks on their composite school value added, compared with 43% of schools below the median URM school. More formal empirical research also suggests that increasing the quality of clinical placement schools can be met without sacrificing attempts to place candidates in more diverse clinical settings (Charner-Laird et al., 2017; Ronfeldt, 2012).[35] Moreover, should these clinical evaluation effects translate into inservice differences in teacher quality, novice teachers are far more likely to be serving students from economically disadvantaged families, rural students, and students of color (Gagnon & Mattingly, 2015). Thus, improvements in teacher candidate teaching capacities would be expected to have a differentially positive impacts on these populations of students. However, in practice, eligibility requirements for mentor teachers could affect the distribution of placements if equity is not a priority for preparation programs. More empirical work is needed to understand whether policies that create eligibility requirements for mentors or placement sites affect the equitable distribution of new teachers or teacher candidates, and state policymakers may want to consider student diversity as a component of placement. Still, there is little reason to view the objectives of improving placement quality and diversity as incompatible.

We also examine the independent effects of mentor characteristics that were included in our overall index for placement quality. Using the clinical evaluation ratings of field instructors as a measure of the instructional skills of teacher candidates, we find consistent evidence that the teaching experience and observational ratings of mentor teachers are associated with the

growth in skills of teacher candidates during clinical practice. There is, by contrast, little evidence that the value added of teachers who serve as mentors is related to the development of instructional skills of teacher candidates during their clinical placements. These findings are robust to a variety of specifications and, in particular, appear to be verified by the experimental assignment of teacher candidates to higher- and lower-index clinical placements.

It is worth noting that it is somewhat challenging to put our findings into context as there is so little quantitative evidence about the growth of preservice teacher candidate skills during their clinical placements. Our findings can be juxtaposed against some of the estimated relationships between mentor teacher characteristics and the inservice performance and effectiveness of the teachers they supervised. Given our exploratory findings surrounding the mechanism of the placement effect, our results strongly suggest benefits associated with assignment to more experienced mentor teachers and to mentor teachers with higher observational ratings. This conflicts with research (Goldhaber, Krieg, & Theobald, 2017; Ronfeldt, Brockman, & Campbell, 2018) that finds little evidence of a connection between the teaching experience of mentors and the observational ratings and value-added effectiveness of their mentees. One possible explanation is that the influence of these teacher characteristics on teacher candidates' instructional abilities weakens over time, lessening the importance of mentors as teacher candidates transition to their first job. The findings are also somewhat in conflict with growing evidence (Goldhaber, Krieg, & Theobald, 2020; Ronfeldt, Brockman, & Campbell, 2018) from research on teacher candidates who enter the labor market that the value added of teachers serving as mentors is significantly associated with the value added of the teacher candidates they supervise during clinical practice.

On the other hand, the sole study (Ronfeldt, Brockman, & Campbell, 2018) that links the value added of mentor teachers to the inservice observational ratings of the teacher candidates they supervised finds an insignificant relationship between these measures. It does, however, find a positive relationship between *observational* ratings of teachers serving as mentors and the observational ratings of their mentees who become teachers. Both of these findings are consistent with our findings for clinical practice. The authors go on to speculate that one possible explanation is that observation ratings and value added may measure different aspects of instructional quality and that mentors are teaching mentees the aspects of instructional quality on which they excel most.

It will be important to see whether the findings about placement quality carry forward when teacher candidates enter the labor market. It is possible that teacher candidate skill growth induced by higher-index placements does not translate well into early career instructional practices or that, after leaving supportive environments, these teacher candidates do not continue to outperform teacher candidates from lower-index placements. Nonetheless, we believe

the findings concerning the effect of placement quality on clinical practice are promising, that is, a plausibly causal relationship between the quality of clinical placements and improved clinical practice exists. Moreover, the mentor characteristics that appear to matter for teacher candidate development, for example, experience and observational ratings, are often tracked by state departments of education and sometimes publicly available making targeted recruitment possible.

More generally, our findings point to teacher education, and clinical placements in particular, as leverage points for growing the instructional skills of teacher candidates. Importantly, this is causally rigorous evidence that easily identified characteristics of mentor teachers can influence the skill development of teachers before they enter the workforce. There is surprisingly little quantitative evidence on how mentor teachers and clinical sites should be recruited so as to develop the skills of prospective teachers. Moreover, there is a lack of quantitative evidence on how clinical placements and mentoring should be structured to further the skill development of teacher candidates. This is surprising in light of the fact that clinical practice is widely regarded as foundational to the development of teaching capacities. While we provide some of the first evidence in both of these regards, we argue that more experiments about teacher education should be carried out in order to continue deepening our understanding of the ways that clinical placements and mentors influence the future effectiveness of teacher candidates.

ORCID iDs

Dan Goldhaber https://orcid.org/0000-0003-4260-4040
James Cowan https://orcid.org/0000-0002-3131-1077
Trevor Gratz https://orcid.org/0000-0002-0050-6496

Supplemental Material

Supplemental material for this article is available online.

Notes

We appreciate the generous financial support that was provided for this research by the Bill and Melinda Gates Foundation. Emanuele Bardelli from the Institute of Education Sciences (IES), U.S. Department of Education (PR/Award R305B150012). This project would not have been possible without the participation of our Teacher Education Program partners in Florida and Tennessee, and the data provided by the Florida and Tennessee Departments of Education, and to key individuals at anonymous teacher education programs for data provision as well as their help on the ground in carrying out this experiment. Please note that the views expressed are those of the authors and do not necessarily reflect those of this study's sponsor, the Florida and Tennessee Departments of Education, participating teacher education programs, or the institutions to which the authors are affiliated. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

[1]In most TEPs, student teachers are required to receive feedback from mentor teachers—as well as their counterparts on the university side, field instructors—in the form of formal evaluations ("clinical evaluations"). Clinical evaluations are, for instance, required by the Council for the Accreditation of Educator Preparation (CAEP), a nationally recognized teacher education organization that accredit TEPs (CAEP, n.d.).

[2]In the other experiment, the "Feedback Initiative," teacher candidates were randomly assigned to a treatment group which received more and better contextualized feedback about their student teaching. Work on the Feedback Initiative is ongoing.

[3]We argue that this level of randomization is necessary to distinguish the effects of TEP features from the effects that may be related to the schools and districts that tend to be served by teacher candidates from particular programs; see Ronfeldt, Bardelli, et al. (2020) for a more in depth discussion of this point.

[4]The only deviation from normal recruitment practices is that we asked TEPs to overrecruit in each of these blocks (usually one to four teachers per block) since some placements inevitably fall through and we wanted to ensure that replacement mentors could be drawn from an assigned list if possible. Although we have no way of identifying which mentors would have been among those overrecruited, TEPs used their existing networks to identify potential mentors, and we have no reason to believe this would have significantly affected the placement pool. We then linked all recruited mentor teachers to administrative and evaluation data on them and to the schools in which they worked, in order to construct the index for higher and lower quality placements. In short, because we relied on variation in (hypothesized) placement quality within the set of status quo placements, we do not believe the experiment would have had much if any effect on the counterfactual distribution of candidate outcomes and as such may ameliorate ethical concerns of randomizing teacher candidates to better or worse placements.

[5]For more information see https://www.tn.gov/content/dam/tn/education/data/tvaas/tvaas_technical_documentation_2017.pdf and http://www.fldoe.org/teaching/performance-evaluation/.

[6]Districts may produce their own rubric and observational model and apply for approval by the department of education. For more information on TEAM see https://team-tn.org/evaluation/teacher-evaluation/.

[7]In Florida, we use an estimate of the school average teacher experience in place of teacher retention, as retention was unavailable.

[8]But, when examining edTPA passing status, the authors find the opposite: a statistically significant relationship between passing and future reading value added and no statistically significant relationship between passing the edTPA and math value added (these seeming contradictory findings are related to nonlinearities across the two outcomes in the relationship between edTPA performance and value added).

[9]The balance test is a joint test of the treatment coefficients from different models on individual teacher candidate characteristics, that is, race, gender, and GPA. We stratify our balance test by Placement block.

[10]This omnibus test requires field instructors to be paired with teacher candidates randomized to both higher- and lower-index placements. Additionally, when stratifying by randomization block this test requires field instructors to be paired with teacher candidates from at least two randomization blocks. A total of 13 field instructors were paired with 20 teachers that were either the same treatment statuses or randomization block and thus dropped from this test.

[11]In the case of teacher experience, the average lower-index mentor teacher has 7.2 years of experience, and the average higher-index mentor teacher has 15.0 years of experience.

[12]The average mentor teacher in the sample has over 11 years of teaching experience.

[13]One program had a 5-point rating scale but used software to record observations that permitted decimal ratings. Some field instructors used decimal ratings (usually in half-point increments). Unless otherwise noted, we leave these ratings as recorded by the evaluator.

[14]And, for example, the Florida TEP's domains ("Instructional Delivery," "Learning Environment," and "Assessment") are closely aligned to three of the foundational principles ("Instructional Design and Lesson Planning," "Learning Environment," & "Assessment") in Florida's Educator Accomplished Practices (Florida Administrative Code, 2011). Source

http://www.fldoe.org/teaching/performance-evaluation/fl-state-models-of-evaluation-system.stml.

[15]Results are similar when we omit the TEP from Florida and available on request.

[16]The Tennessee TEP had a yearlong clinical experience; the Florida TEP had a clinical experience spanning two consecutive semesters, although not necessarily within a single academic or calendar year.

[17]They were, for instance, just 0.5% of first time clinical evaluation ratings.

[18]Recall that the Florida program uses a rating scale of "Not Using," "Beginning," "Developing," "Applying," and "Innovating." However, candidates from the Tennessee TEP constitute the majority of the sample, and as such we have opted to display the evaluation rubric categories from Tennessee.

[19]Given the relatively high marks earned by the majority of teacher candidates, there is some concern that this does not leave enough identifying variation left over to identify a treatment effect. However, recent work by Kraft et al. (2020) find that higher clinical evaluations are associated with higher value added to student achievement throughout the full range of the evaluation range. The authors conclude that "they [clinical evaluation ratings] contain important information about teacher effectiveness despite limited variation in final ratings." In addition, in-service observational ratings often exhibit similar distributions, but are nonetheless predictive of teacher contributions to student test scores (Cowan, Goldhaber, & Theobald, 2020).

[20]The model includes field instructor, teacher candidate, and program-item random effects.

[21]All evaluations included in our models were submitted by university faculty, that is, field instructors. Using evaluations from mentor teachers is problematic because three quarters of the index used to randomize teacher candidates to placement lists came from mentor teacher attributes. That is, mentor teachers are systematically different across placement lists, and therefore could provide systematically different evaluations.

[22]In alternative specifications we model Equation (1) as an ordinal logit model, a graded response model, and replace the ordinal position of the evaluations with the number of weeks since the start of the clinical placement and find similar results. Results are available on request.

[23]Although we relied on a two-by-two factorial design that should balance treatment status in the feedback initiative across placement lists, we additionally control for feedback treatment in Equation (1).

[24]With the exception of some small blocks during the first year of the experiment, all teacher candidates had a probability of 0.5 of receiving a higher-index placement.

[25]For instance, one TEP defines the "Environment" on their evaluation rubric as setting expectations, managing student behavior, creating a welcoming and organized classroom, and establishing a respectful culture. All of these attributes of a classroom environment may be associated with student achievement and/or the perception of teacher candidates' abilities.

[26]Vagi et al. (2019) find that teacher candidates who enter TEPs with higher grade point averages tend to show greater gains on a clinical evaluation rubric rating during the course of their clinical placements.

[27]There are advantages and disadvantages to these two sets of models. Controlling for a lagged clinical evaluation not only allows for controlling for potential time-variant sources of bias but also conditions on a postrandomization variable. This is a potentially conservative approach as it nets out any potential effect from the prior intervention. In an alternative specification, we difference the first and third ratings by candidate and regress the difference on the higher-index placement list. This nets out gains made before the first evaluation, but not afterward. Results are available on request. On the other hand, the inclusion of a teacher candidate fixed effect does not account for time variant sources of bias.

[28]A small number of observations are missing this block average in which case imputation is done at the TEP-cohort level.

[29]In Supplemental Appendix G (available in the online version of the journal), we present a domain analysis where we interact the higher-index indicator with the three clinical evaluation domains. Teacher candidates on the high-index list grow the most in the "Instruction" domain.

[30]The clinical ratings are estimated to increase by about 37% of a standard deviation for the second observation relative to the first and by 55% of a standard deviation for the third relative to the first. This is broadly consistent with the only other quantitative study that documents changes in clinical evaluation ratings during a clinical placement: Vagi et al. (2019) find that teacher candidates improve (over a yearlong clinical placement) by about 44% of a standard deviation from their first to third observation.

[31]The variance of the field instructor fixed effects, which does not account for sampling error, is about 1.8 times as large.

[32]We also run models controlling for specific mentor characteristics and simultaneously the suite of school characteristics. Results are quantitatively similar and available on request.

[33]Note that it is still possible to find "treatment effects" in models that include the components that make up the index determining treatment status. This could occur, for instance, because of nonlinearities in the relationship between the index components and the outcome or interactions between the different components. For instance, if mentor teachers' experience has little effect on candidates beyond the first few years of teaching, the higher-index placement list may improve candidate outcomes by reducing exposure to novice mentors. The random assignment effect may still appear positive in models that control for mentor experience in a linear specification.

[34]An exception to this is Louisiana, which, under the Believe and Prepare program, now requires among other things, a yearlong clinical placement, specialized training for mentor teachers and mentor certification, and selection of mentors using student growth data.

[35]In this study higher-index placements tended to be in wealthier and less African American schools, but not by large margins. For instance, the average school on the higher-index placement list was 5.4% African American compared with 6.5% on the lower-index list. Similarly, 35.3% and 39.5% of students linked to higher- and lower-index placement lists, respectively, were eligible for free or reduced-priced lunch. These findings are consistent with the descriptive analysis of low-turnover schools in Ronfeldt (2012), who shows that many schools serving low-income and non-White students also have strong professional climates. The strength of these professional environments may be largely dependent on support and expectation setting by principals (Charner-Laird et al., 2017).

References

American Association of Colleges of Teacher Education. (2010). *AACTE endorses NCATE Blue Ribbon Panel Report on Clinical Preparation*. https://aacte.org/2010/11/aacte-endorses-ncate-blue-ribbon-panel-report-on-clinical-preparation/

American Association of Colleges of Teacher Education. (2018). *A pivot towards clinical practice, its lexicon, and the renewal of educator preparation: A report of the AACTE Clinical Practice Commission*. http://www.nysed.gov/common/nysed/files/cpc-aactecpcreport.pdf

Anderson, L. M., & Stillman, J. A. (2013). Student teaching's contribution to preservice teacher development: A review of research focused on the preparation of teachers for urban and high-needs contexts. *Review of Educational Research*, *83*(1), 3–69. https://doi.org/10.3102/0034654312468619

Bartanen, B., & Kwok, A. (2020). *Pre-service teacher quality and workforce entry* (EdWorkingPaper No. 20-223). Annenberg Institute at Brown University. https://www.edworkingpapers.com/sites/default/files/ai20-223.pdf

Bastian, K. C., Lys, D., & Pan, Y. (2018). A framework for improvement: Analyzing performance-assessment scores for evidence-based teacher preparation program reforms. *Journal of Teacher Education*, *69*(5), 448–462. https://doi.org/10.1177/0022487118755700

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

Borko, H., & Mayfield, V. (1995). The roles of the cooperating teacher and university supervisor in learning to teach. *Teaching and Teacher Education*, *11*(5), 501–518. https://doi.org/10.1016/0742-051X(95)00008-8

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440. https://doi.org/10.3102/0162373709353129

Burgess, S., Rawal, S., & Taylor, E. S. (2021). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. *Journal of Labor Economics*, *39*(4), 1155–1186. https://doi.org/10.1086/712997

Campbell, S. (2014). *Quality teachers wanted: An examination of standards-based evaluation systems and school staffing practices in North Carolina middle schools* [Doctoral dissertation, University of North Carolina at Chapel Hill Graduate School]. https://cdr.lib.unc.edu/concern/dissertations/td96k3484

Campbell, S. L. (2020). Ratings in Black and White: A quantcrit examination of race and gender in teacher evaluation reform. *Race Ethnicity and Education*. Advance online publication. https://doi.org/10.1080/13613324.2020.1842345

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*, *55*(6), 1233–1267. https://doi.org/10.3102/0002831218776216

Charner-Laird, M., Ng, M., Johnson, S. M., Kraft, M. A., Papay, J. P., & Reinhorn, S. K. (2017). Gauging goodness of fit: Teachers' responses to their instructional teams in high-poverty schools. *American Journal of Education*, *123*(4), 553–584. https://doi.org/10.1086/692663

Chen, B., Cowan, J., Goldhaber, D., & Theobald, R. (2019). *From the clinical experience to the classroom: Assessing the predictive validity of the Massachusetts candidate assessment of performance* (CALDER Working Paper No. 223-1019). https://caldercenter.org/publications/clinical-experience-classroom-assessing-predictive-validity-massachusetts-candidate

Clarke, A., Triggs, V., & Nielsen, W. (2014). Cooperating teacher participation in teacher education: A review of the literature. *Review of Educational Research*, *84*(2), 163–202. https://doi.org/10.3102/0034654313499618

Connor, K., & Killmer, N. (1995, October 11-14). *Evaluation of cooperating teacher effectiveness* [Paper presentation]. Annual Meeting of the Midwest Educational Research Association, Chicago, IL, United States. https://eric.ed.gov/?id=ED394950

Council for the Accreditation of Educator Preparation. (n.d.). *Standard 2: Clinical Partnerships and Practice*. http://caepnet.org/standards/2022-itp/standard-2

Cowan, J., Goldhaber, D., Jin, Z., & Theobald, R. (2020). *Teacher licensure tests: Barrier or predictive tool?* (CALDER Working Paper No. 245-1020). https://caldercenter.org/sites/default/files/WP%20245-1020_0.pdf

Cowan, J., Goldhaber, D., & Theobald, T. (2020). *Performance evaluations as a measure of teacher effectiveness when standards differ: Accounting for variation across classrooms, schools, and districts* (CALDER Working Paper No. 197-0618-2). https://caldercenter.org/publications/performance-evaluations-measure-teacher-effectiveness-when-standards-differ-accounting

Cuenca, A., Schmeichel, M., Butler, B. M., Dinkelman, T., & Nichols, J. R., Jr. (2011). Creating a "third space" in student teaching: Implications for the university supervisor's status as outsider. *Teaching and Teacher Education*, *27*(7), 1068–1077. https://doi.org/10.1016/j.tate.2011.05.003

Florida Administrative Code. (2011). *6A-5.065: The Educator Accomplished Practices*. https://www.flrules.org/gateway/ruleno.asp?id=6A-5.065

Florida Department of Education. (2019). *The Florida Educator Accomplished Practices (FEAPs)*. http://www.fldoe.org/teaching/professional-dev/the-fl-educator-accomplished-practices.stml

Gagnon, D. J., & Mattingly, M. J. (2015). Rates of beginning teachers: Examining one indicator of school quality in an equity context. *Journal of Educational Research*, *108*(3), 226–235. https://doi.org/10.1080/00220671.2013.878300

Ganser, T. (2002). How teachers compare the roles of cooperating teacher and mentor. *Educational Forum*, *66*(4), 380–385. https://doi.org/10.1080/00131720208984858

Gareis, C. R., & Grant, L. W. (2014). The efficacy of training cooperating teachers. *Teaching and Teacher Education*, *39*(April), 77–88. https://doi.org/10.1016/j.tate.2013.12.007

Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2021). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, *58*(1), 3–31. https://doi.org/10.3102/0002831219890608

Glenn, W. J. (2006). Model versus mentor: Defining the necessary qualities of the effective cooperating teacher. *Teacher Education Quarterly*, *33*(1), 85–95. https://eric.ed.gov/?id=EJ795211

Goldhaber, D., Cowan, J., & Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education*, *68*(4), 377–393. https://doi.org/10.1177/0022487117702582

Goldhaber, D., Krieg, J. M., & Theobald, R. (2017). Does the match matter? Exploring whether student teaching experiences affect teacher effectiveness. *American Educational Research Journal*, *54*(2), 325–359. https://doi.org/10.3102/0002831217690516

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

Goldhaber, D., Krieg, J. M., Naito, N., & Theobald, R. (2020). Making the most of student teaching: The importance of mentors and scope for change. *Education Finance and Policy*, *15*(3), 581–591. https://doi.org/10.1162/edfp_a_00305

Goldhaber, D., Quince, V., & Theobald, R. (2019). Teacher quality gaps in US public schools: Trends, sources, and implications. *Phi Delta Kappan*, *100*(8), 14–19. https://doi.org/10.1177/0031721719846883

Graham, B. (2006). Conditions for successful field experiences: Perceptions of CTs. *Teaching and Teacher Education*, *22*(8), 1118–1129. https://doi.org/10.1016/j.tate.2006.07.007

Greenberg, J., Pomerance, L., & Walsh, K. (2011). *Student teaching in the United States*. National Council on Teacher Quality. https://www.nctq.org/dmsView/Student_Teaching_United_States_NCTQ_Report

Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, *40*(June), 183–204. https://doi.org/10.1016/j.econedurev.2014.03.002

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Bill and Melinda Gates Foundation.

Hoffman, J. V., Wetzel, M. M., Maloch, B., Greeter, E., Taylor, L., DeJulio, S., & Vlach, S. K. (2015). What can we learn from studying the coaching interactions between CTs and preservice teachers? A literature review. *Teaching and Teacher Education*, *52*(November), 99–112. https://doi.org/10.1016/j.tate.2015.09.004

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587–613. https://doi.org/10.1353/jhr.2011.0010

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

Krieg, J. M., Theobald, R., & Goldhaber, D. (2016). A foot in the door: Exploring the role of student teaching assignments in teachers' initial job placements. *Educational Evaluation and Policy Analysis*, *38*(2), 364–388. https://doi.org/10.3102/0162373716630739

Mansfield, R. (2015). Teacher quality and student inequality. *Journal of Labor Economics*, *33*(3), 751–788. https://doi.org/10.1086/679683

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*, *71*(1), 41–62. https://doi.org/10.1177/0022487118791992

National Council for Accreditation of Teacher Education. (2010). *Transforming teacher education through clinical practice: A national strategy to prepare effective teachers* (Report of the Blue Ribbon Panel on clinical preparation and partnerships for improved student learning). http://www.highered.nysed.gov/pdf/NCATECR.pdf

National Council on Teacher Quality. (2017). *A closer look at student teaching: Undergraduate secondary programs*. https://www.nctq.org/dmsView/US_2017_ST_Findings

National Council on Teacher Quality. (2020). *Student teaching/clinical practice national results: State teacher policy database* [Data set]. https://www.nctq.org/yearbook/national/Student-Teaching/Clinical-Practice–81

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

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, *12*(1), 359–388. https://doi.org/10.1257/pol.20170709

Rice, J. K. (2010). *The impact of teacher experience: Examining the evidence and policy implications* (CALDER Brief No. 11). https://www.urban.org/sites/default/files/publication/33321/1001455-The-Impact-of-Teacher-Experience.PDF

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

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. https://doi.org/10.3102/0162373711420865

Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education*, *66*(4), 304–320. https://doi.org/10.1177/0022487115592463

Ronfeldt, M., Bardelli, E., Truwit, M., Mullman, H., Schaaf, K., & Baker, J. C. (2020a). 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., & Campbell, S. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher*, *47*(7), 405–418. https://doi.org/10.3102/0013189X18782906

Ronfeldt, M., Matsko, K. K., Nolan, H. G., & Reininger, M. (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). https://cepa.stanford.edu/content/who-knows-if-our-teachers-are-prepared-three-different-perspectives-graduates-instructional-readiness-and-features-preservice-preparation-predict-them

Ronfeldt, M., Matsko, K. K., Nolan, H. G., & Reininger, M. (2020). 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

St. John, E., Goldhaber, D., Krieg, J., & Theobald, R. (2018). *How the match gets made: Exploring student teacher placements across teacher education programs, districts, and schools* (CALDER Working Paper No. 111018). https://caldercenter.org/sites/default/files/CALDER%20WP%20204-1018-1.pdf

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317. https://doi.org/10.3102/0162373715616249

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Steinberg, M. P., & Sartain, L. (2020). What explains the race gap in teacher performance ratings? Evidence from Chicago public schools. *Educational Evaluation and Policy Analysis*, *43*(1), 60–82. https://doi.org/10.3102/0162373720970204

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

Vagi, R., Pivovarova, M., & Barnard, W. (2019). Dynamics of preservice teacher quality. *Teaching and Teacher Education*, *85*(October), 13–23. https://doi.org/10.1016/j.tate.2019.06.005

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations*. Brown Center on Education Policy, Brookings Institute. https://www.brookings.edu/wp-content/uploads/2016/06/Evaluating-Teachers-with-Classroom-Observations.pdf

Zeichner, K. M. (2009). *Teacher education and the struggle for social justice*. Routledge. https://doi.org/10.4324/9780203878767

Manuscript received June 5, 2020
Final revision received November 1, 2021
Accepted November 15, 2021