



NATIONAL
CENTER *for* ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington



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

**Dan Goldhaber, Matt
Ronfeldt, James
Cowan, Trevor Gratz,
Emanuele Bardelli, Matt
Truwit, Hannah
Mullman**

Contents

Contents	i
Acknowledgments.....	ii
Abstract	iii
Introduction.....	1
Background on Clinical Practice and the Improving Student Teaching Initiative	2
Data and Sample Description.....	7
Analytic Approach	13
Results.....	18
Discussion and Conclusions	26
References.....	30
Figures and Tables	36
Appendix A: Evaluation Ratings by Instructor Type	44
Appendix B: Mentor Teacher Evaluations by Mentor Characteristics	45
Appendix C: Effects of Mentor Teacher and School Characteristics on Field Instructor Evaluations.....	46
Appendix D: Quasi-Experimental Mentor Teacher Characteristic Models.....	50
Appendix E: Clinical Evaluation Level Analyses	54
Appendix F: Placement Characteristics Interacted with the Treatment.....	56
Appendix G: Treatment Models with Block Fixed Effects	57

Acknowledgments

We appreciate the generous financial support that was provided for this research by the Bill and Melinda Gates Foundation. Emanuele Bardelli and Hannah Mullman received pre-doctoral support 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.

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007
202-403-5796 • www.caldercenter.org

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

Dan Goldhaber, Matt Ronfeldt, James Cowan, Trevor Gratz, Emanuele Bardelli, Matt Truwit, Hannah Mullman

CALDER Working Paper No. 239-0620

June 2020

Abstract

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 university instructors. The findings suggest that improving clinical placement procedures can improve the teaching quality of candidates.

Introduction

Teacher preparation has emerged as a high profile area of focus for reforms of policy and practice (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, NCATE, 2010); indeed, it has been identified as “a key component—even ‘the most important’ component—of pre-service 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 student teachers receive high quality mentoring during their internships (AACTE, 2010; NCATE, 2010).

The role of clinical placements are potentially profound as the guidance and evaluations provided by mentor teachers to student teachers 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 internship placement, as measured by mentor effectiveness or school quality, and their future performance in the classroom (Goldhaber, et al., 2018; Ronfeldt, 2015; Ronfeldt et al., 2018a, Ronfeldt et al., 2018c). 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 a within teacher education program (TEP) randomized control trial to assess the causal relationship between the quality of internship placements and the development of 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. During the internship university field instructors and mentor teachers provided regular assessments of candidates’ teaching practice using observational rubrics aligned with

inservice teacher evaluation frameworks. Prior research has found 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 internships (Vagi, Pivovarova, & Barnard, 2019). We exploit the randomization of teacher candidates to either high or low quality placements to study the effects of placement quality on the 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 section) clinical placements saw greater increases, by 0.06 standard deviations, in ratings over the course of their internship. This represents approximately 12% of the total increase in clinical evaluation scores from the beginning of the placement until the end of the placement. Although non-experimental estimates of the contributions of individual school and mentor characteristics are imprecise, they suggest that mentor teacher experience and, less certainly, school turnover and observational ratings positively influence candidates’ observed performance. These findings on teacher candidates are consistent with prior observational studies of inservice teacher outcomes suggesting that candidates learn measurable teaching skills from higher quality clinical placements (Goldhaber et al., 2020a; Ronfeldt et al., 2018a; Ronfeldt et al., 2018b; Ronfeldt et al., 2018c).

Background on Clinical Practice and the Improving Student Teaching Initiative

Background on Mentor Teachers and Clinical Evaluations

Numerous qualitative studies (Clarke et al., 2014; Ganser, 2002; Graham, 2006; Hoffman et al., 2015; Zeichner, 2009) document the myriad roles mentor teachers 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 via formalized clinical evaluations.

A growing body of research finds evidence that mentor quality matters for the development of teacher candidates. Ronfeldt et al. (2018a, 2018c) 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 et al. (2018a) and Goldhaber et al. (2018) find that the effectiveness of mentor teachers (measured by value added) is significantly associated with the later effectiveness of their mentees who enter teaching. And in an experimental evaluation of the effects of assigning candidates to more effective mentors, Ronfeldt et al. (2018b) find that candidates placed with more instructionally effective mentors also report receiving higher quality and more frequent coaching and feedback.

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

Clinical evaluations are likely the first feedback that teacher candidates receive based on practice in authentic settings. Evidence from inservice teachers suggests that structured feedback

based on evaluations can improve instructional practice (Burgess, Rawall, & Taylor, 2019; Garet et al., 2017; Papay, Taylor, Tyler, & Laski, 2020; Steinberg, & Sartain, 2015; Taylor & Tyler, 2012). Taylor & Tyler (2012) find that students taught by inservice teachers who are evaluated over the course of a school year, score 0.10 standard deviations higher on standardized math tests than students taught by the same teacher before the evaluation. Moreover, results from a recent randomized control trial of inservice teachers demonstrate that low stakes peer performance evaluations increase educator effectiveness. Students taught by teachers in the treatment group (teachers receiving peer evaluations) scored 0.07 standard deviations higher on standardized math and reading tests than those taught by teachers in the control group (Burgess, Rawall, & Taylor, 2019). Learning from highly effective teachers via peer evaluations may be especially important for educators (Papay, Taylor, Tyler, & Laski, 2020; Taylor & Tyler, 2012).

Yet there is evidence that the degree to which mentors devote time and energy to providing feedback, along with the nature of the feedback, varies a good deal among mentors (Borko & Mayfield, 1995; Clark et al., 2014; Mastko et al., 2018). 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 evaluate and provide student teachers with feedback (NCTQ, 2016; NCTQ, 2017). However, there are no quantitative studies that link mentor characteristics to the clinical evaluations that teacher candidates receive.

Importantly, student teaching is typically 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 local schools/districts in which clinical practice takes place (Cuenca, Schmeichel, Butler, Dinkelman, & Nichols, 2011). Field instructors are theorized to help student teachers 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 referred to as “university supervisors”) provide “an objective, third-party viewpoint in the student teaching triad” (Gareis and Grant, 2014). And it is the clinical evaluations of field instructors that we utilize as our measures of student teacher development and performance.

Background on the Improving Student Teaching Initiative (ISTI)

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 student teaching placements on teacher candidates’ perceptions of their preparedness to teach, preservice 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 internships, as judged based on an index of mentor teacher and internship school characteristics (described in greater detail below).²

The implementation of the Placement Initiative included two key features: within TEP randomization³ and the collection of clinical practice evaluations of teacher candidates by mentor teachers and field instructors. The Placement Initiative was designed to isolate the causal effect of placement quality. Estimating the relationship between internship characteristics and teacher candidate outcomes without random assignment could be biased by the matching of candidates and internship sites. Krieg, Theobald, and Goldhaber (2016), for instance, find evidence that student teachers who have higher basic skills licensure tests are more likely to be matched to mentor teachers who also have higher basic skills licensure tests.

To account for the concern about the matching process, we randomized candidates to one of two lists of placements. We first asked programs to assemble a list of potential internship placements in advance. We then matched the host schools and mentor teachers to state administrative databases and ranked placements based on three attributes of the mentor teachers (experience, observational performance ratings, and a state-supplied measure of teacher value added) and two attributes of schools (school value added and turnover rates).⁴ The attributes of the index came from two different states (Florida and Tennessee). Consequently, the measures we used differed slightly in each; we combined these measures into single teacher and school aggregates.⁵ We combined the teacher and school aggregates into an overall internship quality index, applying a weight of 0.75 to the mentor teacher aggregate and 0.25 to the school-level aggregate.

Teacher candidates were categorized into blocks to accommodate geographic preferences (district or county), subject area (e.g. math, social studies, self-contained) and grade level (e.g., elementary grades, high school), and programs were encouraged to over-recruit potential mentor teachers within each block. We divided internship placements within each block into “higher” and “lower” quality lists based on whether the index was above or below the median value in the block. In this paper, we refer to the higher list as being higher quality and having more “instructionally effective” teachers. 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 higher or lower values of the placement index, which was constructed from measures identified in the literature that are associated with “better” student outcomes. After identifying the two potential placement lists for each block, we randomly assigned teacher candidates to either higher or lower quality placements and returned the two sets of lists to the programs. The

programs then assigned candidates to a 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 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, 2016-17 and 2017-18, in two teacher education programs (one program in Florida and one in Tennessee); in total it included 398 teacher candidates (95% of whom attended the Tennessee TEP). Of these teacher candidates, 8 candidates randomized to the high index list and 12 randomized to the low index list either were delayed in the program, did not meet residency requirements, or changed majors, resulting in an analytic sample of 369 candidates. Importantly, these candidates were dropped from the experiment before beginning their teaching residency and their attrition is not likely to be caused by exposure to higher quality placements. In addition, 3 low index and 6 high index candidates could not be matched to evaluation data obtained during the experiment.

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 datasets. These datasets include information on official state value-added measures, observational performance measures, teacher experience, school value-added measures, and teacher retention.

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).⁶ Teacher performance ratings in Tennessee are conducted multiple times throughout the year, and the evaluative rubric is provided by the Tennessee Educator Acceleration Model (TEAM).⁷ Performance ratings in Florida are developed by districts in accordance with Florida’s “Educator Accomplished Practices” (FAC 6A-5.065, 2011), and we utilize the teachers’ final district performance rating on a four-point scale. These same administrative records were used to construct average school value-added and school level measures of teacher retention.⁸ As described in the ISTI Background section, we employ this data to construct two lists of high and low quality teacher candidate clinical placements and randomize candidates to these placement lists.

The evaluation data used as focal outcomes for this study come from the clinical evaluations of mentor teachers and field instructors, and was 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 evaluation scores each term. We linked this data to random assignment information we had compiled before each placement began.

Table 1 reports the number of teacher candidates participating in the Placement Initiative, as well as summary statistics by treatment status (Panel A). The differences between the teacher candidate characteristics in the high- and low-quality placements are generally quite small, suggesting that the randomization was successful. But we formally test this using an omnibus balance test between high and low index pre-experiment teacher candidate characteristics.⁹ The

p-value of 0.262 reveals no significant differences between teacher candidates by randomization status.

While teacher candidates were randomized to high or low indexed mentor teachers, one might still be concerned that there might be systematic matching of field instructors to teacher candidate-mentor teacher dyads based upon the quality of the mentor teacher. For instance, if field instructors who are “tough graders” when it comes to clinical observation ratings are matched to more experienced mentor teachers, we would see a downwardly biased relationship between mentor experience and field instructor ratings. To address this concern, in panel C of Table 1 we conduct an additional omnibus balance test of field instructors using field instructor indicators by the treatment status of their teacher candidates. Results from this test (p-value = 0.49) indicate that this type of sorting did not occur.¹⁰

Panel B reports the average mentor teacher characteristics by Placement Status, and as designed, we see significant differences between the mentor teacher characteristics of teacher candidates on the low index placement list compared to the high index placement list. For instance, teacher candidates assigned to the high index placement were placed with mentors whose average values for the value-added measure, observational performance ratings, and experience exceeded the average values of the lower index placement mentors by 0.7 standard deviations or more for each measure.¹¹ The differences in school characteristics between the high and low 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). 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 high index placement for its geographic-grade-subject randomization block. The mean difference between the high and low 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 teachers. In particular, the correlation between standardized mentor teacher observation ratings and experience is 0.25, and the correlation between mentor teacher observation ratings and value added is 0.28. The correlation between mentor teacher value added and experience was negative (-0.05) but not significant; this is not terribly surprising given the small sample of mentor teachers and the fact that mentor teachers were 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 a score 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 five-point Likert 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, innovating). A

teacher candidate receiving the highest evaluation score 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 “Applying [teaching skill] to classroom practice” (in Florida) and “Above Expectations” in Tennessee. Whereas the bottom two categories of the scale used the terms “[teaching skill] Not Used” and “Beginning” (in Florida) and “Significantly Below Expectations” and “Below Expectations” in Tennessee.

Both states are working to align expectations of teacher candidates with those of inservice teachers. As such, the clinical evaluation rubric for student teachers is aligned to the evaluation rubric used for inservice teachers. Thus, 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 (FDOE).¹⁴ In Tennessee the TEP evaluation rubric used is the same as the state’s inservice performance evaluation rubric, the TEAM statewide observational rubric. Recall that we have evaluation ratings of mentor teachers as part of their inservice evaluations, as well as ratings by both field instructors and mentor teachers of student teachers. To distinguish between these, we henceforth refer to ratings of mentor teachers as “inservice evaluation ratings” and of student teachers as “clinical evaluation ratings”.

It is the clinical evaluation ratings of the field instructors that are the focal outcomes of this study. In particular, in the analyses we describe below, we focus on whether 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 1. In total we have records for

20,483 ratings spread across 1,126 observations, with teacher candidates from the much larger Tennessee TEP making up 95% of the sample.

On average we have 2.9 field instructor ratings per teacher candidate, with the first evaluation taking place six 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 year long internship.¹⁵

In Figure 2 we provide 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 score 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.^{16, 17} 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 observation, over 70% of teacher candidate-items are rated “Above” or “Significantly Above” expectations, while less than 3% are “Below” or “Significantly Below” expectations. 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, and this is before they are actually hired as teachers. On the other hand, field supervisors and mentor teachers may alter their standards for preservice teachers; for example, their criteria for “meets expectations” may be more lenient for preservice 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 Appendix A, 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),¹⁸ 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 high or low index placement lists, outlined in the Background. We implement an intent-to-treat analysis of the effect of placement quality on evaluation scores with an indicator for assigned placement list, as in Equation (1):

$$Y_{iltjp} = \beta_1 Treat_i + \beta_2 Order_{itjp} + \beta_3 Treat_i * Order_{itjp} + \alpha_{lp} + \theta_j + \omega_i + \epsilon_{iltjp} \quad (1)$$

In Eq. (1), i denotes the individual, l denotes the item, t denotes the timing of the observation (the ordinal position of the rating), j denotes the field instructor, and p denotes the TEP. $Treat_i$ is a binary indicator for whether or not a teacher candidate was randomized to the high index placement list. $Order_{itjp}$ is the ordinal position (i.e. by date) of the rating within candidate-item cells.¹⁹ Equation (1) includes item-by-program (α_{lp}), teacher candidate (ω_i), and field instructor random effects (θ_j).^{20,21}

We hypothesize that exposure to higher quality clinical placements may have a cumulative effect on teacher effectiveness. For instance, Gareis and Grant (2014) found that teacher candidates of more highly trained mentor teachers had faster growth in their clinical evaluation scores over the course of their clinical placement. We investigate this possibility by including in some specifications of Equation (1) an interaction term between $Treat_i$ and $Order_{itjp}$.

Randomization of candidates to placement lists ensures that candidate effectiveness is not correlated with treatment status, which would bias estimates of the coefficients in Eq. (1). However, even with random assignment of candidates to placement lists there are still three primary challenges to interpreting β_1 as the causal effect of treatment on teaching practice.

First, field instructors may conflate the preexisting classroom environment with the teaching practice of candidates.²² This would be consistent with prior research that has found that teachers tend to receive higher observational ratings when they teach in higher achieving classrooms (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). To the extent that clinical placements from the high 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, $(Y_{il(t-1)jp})$, which is depicted in Eq. (2):

$$Y_{iltjp} = \beta_1 Treat_i + \beta_2 Order_{itjp} + \beta_3 Y_{il(t-1)jp} + \alpha_{lp} + \theta_j + \omega_i + \epsilon_{iltjp} \quad (2)$$

This specification removes time-invariant differences in candidate quality that may be correlated with clinical placements, as well as any potential effects clinical placement quality has on teacher candidates through their first clinical evaluation.²³ It therefore identifies the effect of

high quality clinical placements by comparing the growth rates of candidates assigned to higher or lower quality 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 first field instructor evaluation. This would *not* be the case if field instructors place more weight on the clinical placement quality as the internship 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 internship, which suggests that field instructors likely place more weight on the skills of the teacher candidates throughout the year.

A second concern is that mentor teachers may influence the ratings given by field instructors. For instance, more instructionally effective mentor teachers may provide systematically higher or lower evaluations of their candidates. Although we rely on the clinical evaluations collected by field instructors, it is possible that mentors influence field instructor ratings by communicating their perceptions of student teachers to the field instructors. As the index used to construct the treatment was weighted to reflect mentor teacher characteristics, an estimated treatment effect may partially reflect the evaluation standards of the mentor teachers to whom they have been assigned. Although we cannot test this directly, in the Threats to Causal Interpretation section, we do provide evidence that mentor teachers on the high placement 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 randomized list

placement status and field instructor fixed effects. We do not find systematic differences in ratings across placement lists.

Lastly, 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 quality placement, but they did understand the research design and it is possible they correctly identified the lists. If they, for instance, compensated for less promising placements with better (or more generous rating) field instructors, our estimates would tend to be biased downward. However, this does not appear likely in this case. We show in Table 1 that individual field instructors are balanced between treatment 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 to the same field instructor, but across treatment groups, which alleviates the concern of sorting on field instructor rating standards by treatment status. Results are robust to this change and are available upon request.

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

$$Y_{iltjp} = \beta_1 Index_i + \beta_2 Treat_i + \beta_3 Order_{itjp} + \alpha_{lp} + \theta_j + \omega_i + \epsilon_{iltjp} \quad (3)$$

Equation (3) is similar to Equation (1), but now includes a vector of the mentor teacher and school placement characteristics. These characteristics are included in the index ($Index_i$), and include, depending on the specification, teacher value added, experience, and observational ratings and school-level teacher retention and value added. We add variables from this $Index_i$ vector sequentially to see, what if any, effect each variable has on the estimated treatment effect (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 average²⁴, and include both the indicator and imputed values in the index vector. As with Equation (1) in some specifications we interact $Order_{itjp}$ with the treatment indicator ($Treat_i$).

Finally, as Pritchard (2017) describes, student learning is multidimensional and requires teachers to take multidimensional approaches to enhancing student learning. Distinct categories of knowledge and skill reflecting a multidimensional approach are evaluated by TEPs on the clinical evaluation rubrics and are nested under TEP domains. It may be that certain domains are more amenable to influence by high quality clinical placements than others. We investigate this possibility through the following model, Equation (4):

$$Y_{iltjp} = \beta_1 Domain_{ilp} + \beta_2 Treat_i + \beta_3 Domain_{ilp} * Treat_i + \beta_4 Order_{itjp} + \alpha_{lp} + \theta_j + \omega_i + \epsilon_{iltjp} \quad (4)$$

Equation (4) interacts the treatment indicator by the Program-Item specific domains²⁵ from the clinical evaluation rubrics. In some specifications, we include and interact the prior field instructor's item level clinical evaluation ratings, ($Y_{il(t-1)jp}$), with the domain variable. Both TEPs have in common the domains of "Instruction" and "Environment". The TEP that comprises 95% of the sample has a third domain titled "Planning", but which also has evaluation items targeting "Assessment". As such, we combine the other TEP's third domain. "Assessment", with this "Planning" domain and will from here on out refer to it as the "Planning" domain.

Results

Leveraging the Experiment: Clinical Evaluations and High/Low Quality Placements

In **Table 2**, we report the intent-to-treat estimates using the random assignment of teacher candidates to clinical placement lists (consistent with Equation (1) above). These coefficients describe the relationship between field instructor clinical evaluation ratings and clinical placements. Given the standardization of the clinical evaluations, the coefficient estimates should be interpreted as the estimated standard deviation change in field instructor ratings of student teachers on the high index placement list relative to the clinical evaluations of teacher candidates on the low index placement list.

We begin with a specification (column 1) that just includes a placement indicator (equal to 1 if a candidate is in the high placement group relative to low). While the indicator is not statistically significant, the coefficient estimate suggests that teacher candidates in the high placement group tend to receive higher field instructor clinical ratings throughout their internships. In columns 2 and 4, we allow the effect of the placement to accumulate over the course of the internship. We add interactions between rating order and placement in column 2 and find that the third and final rating that teacher candidates receive is significantly higher, by about 6% of a standard deviation, for those in the high placement relative to low placement. Given that the average change between the 1st 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 evaluation scores. In this interaction model the cumulative effect of the high quality placements should be interpreted as the main effect, “High Index Placement” plus the interaction “High Index Placement * 3rd Evaluation”, which results in a treatment effect by the end of the clinical placement of roughly 11% of a standard deviation.

Consistent with the findings reported in Figure 1, there is clear evidence of increased ratings from the 1st to 3rd evaluation of student teachers.²⁶

In Table 3, we display results for models that use outcome data from the second and third observations and control for the prior field instructor evaluation score. Given the adjustment for prior evaluation score in these models, the coefficient on high index placement should be interpreted as an effect on candidate growth during the internship. If the apparent effects of placement list 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 placement list. Although the coefficients on placement are only statistically significant at the 10% level in most of the models, the coefficients, 0.07 to 0.08 standard deviations on the clinical evaluations, are consistent with and somewhat larger than those shown in Table 2. This suggests that candidates on the high placement list improve their measured teaching skills by more than candidates on the lower list throughout the internship.

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 tough a field instructor is 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 rater effect is about 0.109. Because the ratings are standardized, this means that differences in standards and/or impact of coaching across field instructors account for about 10.9% of the variation in candidate ratings.²⁷ 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 (mentor teacher experience, value-added, and observational rating, and school-level retention, and value-added) on the treatment effect.

In Table 4 we add each of the mentor characteristics to the model (columns 3-8) and school characteristics (columns 9-12).²⁸ We interpret the diminishment of the treatment 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 treatment after controlling for the characteristic in question. For example, if the inclusion of mentor teacher experience results in a substantial decrease in the magnitude of the estimated treatment effect and a positive coefficient on experience, we would interpret that as mentor teacher experience being a vital component of the treatment. If on the other hand, the treatment effect is unchanged, then we would interpret that as mentor teacher experience contributing little, if anything, to an estimated treatment effect.²⁹

The inclusion of mentor teacher value added (columns 3 and 4) actually leads to increases in the estimated treatment effect, while the estimate on value added is negative and marginally statistically significant. However, the estimated cumulative effect of the high placement list

(column 4) is relatively unchanged. A possible explanation of the finding in column 3 is that higher value-added teachers are tougher raters and that this may influence field instructor ratings. We explore this in Appendix B, where we report on models that regress the first mentor rating against mentor characteristics. There is evidence consistent with the notion that higher value-added teachers are tougher graders (the point estimate relating value added to mentor ratings is $-.121$), but the estimated relationship is imprecise. Moreover, after controlling for a field instructors' prior evaluation score, i.e. the growth models, the point estimate for teacher value added is -0.027 or roughly 51% of the magnitude of column 3 and insignificant.³⁰

The inclusion of experience in the model as a covariate significantly diminishes the magnitude of the coefficient on placement in column 5 (from 0.065 to 0.011) and, in particular, the ratings that teacher candidates receive on the first evaluation by field instructors (column 6). This is consistent with the positive and significant coefficient on mentor teacher experience. If we add mentor experience to the model represented by column 1 of Table 3 (the growth models) the coefficient on the treatment falls from 0.068 to 0.046. In either case, the addition of mentor experience to our models reduces the treatment coefficient by 32% to 84% depending on the model.³¹

In columns 7 and 8 we include the effects of the inservice observational ratings of mentor teachers. The coefficient on the high quality placement is significantly attenuated with the inclusion of mentor observational ratings, which is consistent with the positive, but insignificant estimates on the observational ratings (columns 7 and 8). When we include mentor observational ratings, the magnitude of the coefficient on high quality placements is 58% of its magnitude when mentor observational ratings are excluded from the model.

Finally, in columns 9-12 we show the inclusion of school characteristics in the model. The findings on the high quality 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 high quality placement coefficient is reduced by only 9% for the school value-added model and 7% for school retention model. This suggests that these school characteristics explain little of the effect of the placement in influencing the clinical observation ratings, but, 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 treatment indicator. When all mentor and school characteristics from the index are entered simultaneously, the coefficient on treatment is 0.002 with a standard error of 0.058.³²

By including the treatment indicator in Table 4 which is consistent with Equation (3), we are implicitly comparing mentor teacher and school characteristics within placement lists. That is, we are only capturing the within treatment category variation of mentor characteristics. In Appendix C, we report results that omit the experimental treatment status so that the results are based on both within and across treatment group variation in mentor and school characteristics (we also estimate models with and without mean imputed placement characteristics). The findings are broadly consistent with those reported above.

In Appendix D, we further develop a quasi-experimental approach for ascertaining the effects mentor characteristics have on clinical evaluations by comparing the within placement block treatment/control differences in clinical evaluations and mentor 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.

Domain Analyses

In Table 5, we present the coefficient estimates from Equation (4) that investigate the potential differential impacts of high quality placements on different dimensions of teaching practice. The treatment indicator, $Treat_i$, has been interacted with the three domains listed on the clinical evaluation rubrics: “Instruction”, “Environment”, and “Planning”. The base level of the main effect of the domain factor variable is “Instruction”, so coefficients should be interpreted as changes from field instructor clinical evaluations on this domain.

Column (1) reports the domain factor main effects and the domain factors interacted with the treatment indicator (which equals 1 for a high quality placement). The interaction between “High Index Placement” and each of the domains shows that candidates assigned to high-index placements appear to perform better on all domains, though the estimated magnitudes differ across the domains. The coefficient on “High Index Placement * Instruction” identifies the effect of high quality placements on “Instruction” domain items. Relative to teacher candidates in lower quality placements, candidates in high quality placements perform 8% of a standard deviation better on “Instruction” domain items, and this is statistically significant. Similarly, candidates at high quality placements perform 0.031 and 0.052 standard deviations better on “Environment” and “Planning” domain items, respectively, than candidates at low quality placements, but these estimates are not statistically significant. Relative to the “Instruction”

domain teacher candidates score 41% of a standard deviation better on the "Environment" domain items and 10% of a standard deviation worse on "Planning" domain items.

Column (2) reports the same coefficients, but adds a control for a teacher candidate's prior item specific evaluation rating by field instructor and interacts this prior rating by domain. Domain main effect coefficients should still be interpreted relative to the "Instruction" domain, however, coefficients now represent growth from the prior evaluation score. Similar to column (1) teacher candidates, overall, perform better on the "Environment" domain and worse on the "Planning" domain relative to the "Instruction" domain items. Moreover, the main take-away from column (1), that high quality placements improve teacher candidates' performance on the "Instruction" domain items, but not "Environment" or "Planning" domain items, is borne out by these growth models.

Threats to Causal Interpretation

The relationship between placement characteristics and candidate performance appears robust to two of the three threats to identification raised in the Analytic Approach. Candidates assigned higher quality placements develop more quickly over time, and those randomly assigned to the higher placement list also earned higher ratings. We also find little evidence that sorting of mentors or 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 mentors 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 candidate.

Suppose ratings consist of baseline candidate efficacy (θ_i), a rater effect (μ_j), and a time effect (λ_t), and a time-variant effect of mentors on true candidate practice (α_{jt}). Then the observed rating is

$$Y_{ijt} = \theta_i + \mu_j + \lambda_t + \alpha_{jt} + \epsilon_{ijt}$$

A concern about our findings when leveraging the experiment is that mentor teacher ratings in the treatment group could be systematically high given candidate ability; that is,

$$\frac{d}{dIndex} E[\mu_j | Index_i] > 0$$

and that these ratings drive up the field instructor ratings. If mentor teachers and field instructors both evaluate the candidate at the same time, then we can difference the mentor teacher and field instructor evaluations to remove the candidate, time, and treatment effects to obtain

$$Y_{ijt} - Y_{ij't} = \mu_j - \mu_{j'} + \epsilon_{ijt} - \epsilon_{ij't}$$

where j denotes the mentor teacher and j' denotes the field instructor. If we regress these differences on field instructor fixed effects and a treatment indicator, we remove the field instructor rater effects and compare the rater effects for mentors assigned the same field instructor. A positive coefficient on treatment list indicates that higher quality mentors provide systematically higher ratings.

We present the results of these tests in Table 6 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 teachers rate their candidates lower than field instructors (again, this is consistent with the findings presented in Appendix B and discussed with the results that leverage the experiment). 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 (i.e. this result at least suggests that the experimental results are a lower-bound on the effect of higher quality placements). More experienced mentors and those on the high placement lists rate their candidates higher, although the differences are not significant. The sign of the coefficient on mentor teacher observational ratings 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 characteristics, although the confidence intervals cannot rule out substantial differences.

Discussion and Conclusions

To our knowledge, this paper is the first to document whether development of teaching skills while in clinical practice is in part attributable to the quality of the field placement and characteristics of mentor teachers. The results suggest that high quality field placements cause candidates' instructional effectiveness to grow during student teaching, as measured by clinical evaluations. This finding is consistent with prior correlational evidence but provides some of the only causal evidence for the importance of high-quality field placements during initial teacher preparation.

We also examine independent effects of mentor characteristics that were included in our overall index for placement quality. Using the ratings of field instructors as a measure of the instructional skills of student teachers, we find consistent evidence that the teaching experience of mentors is associated with the growth in skills of student teachers during clinical practice. To a lesser (and not statistically significant) degree, the observational performance ratings of mentors also appear to be predictive, but this relationship is imprecisely estimated. 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 internships. These findings are robust to a variety of specifications and, in particular, appear to be verified by the experimental assignment of teacher candidates to high- and low-quality 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 internships. Our findings can be juxtaposed against some of the estimated relationships between mentor 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 a lesser extent mentor teachers with higher observational ratings. This conflicts with research (Goldhaber et al., 2017; Ronfeldt et al., 2018a) that finds little evidence of a connection between the teaching experience of mentors and the observational ratings or value-added effectiveness of their mentees. One possible explanation is that the influence of mentor teachers on teacher candidates instructional abilities decays with time and in-class experience, making the relative importance of mentor teacher quality less as teacher candidates transition to their first job. The findings are also somewhat in conflict with growing evidence (Goldhaber et al., 2020a; Ronfeldt et al., 2018a) 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 student teachers they supervise during clinical practice.

On the other hand, the sole study (Ronfeldt et al., 2018a) that links the value added of mentor teachers to the inservice observational ratings of the teacher candidates they supervised, finds a negative relationship between these measures. It does, however, find a positive relationship between *observational* ratings of 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. In that regard, we believe the findings concerning the effect of placement quality on clinical practice are promising, that is, a causal relationship between the quality of clinical placements and improved clinical practice exists. Given that only about three percent of teachers serve as mentors in any year, there is ample room for changing mentor assignments and, subsequently, improving teacher candidate preparation (Goldhaber et al., 2019). Moreover, particular mentor characteristics matter for teacher candidate development, such as experience and evaluation scores, are typically tracked by state departments of education and sometimes publicly available making targeted recruitment possible.

More generally, our findings point to teacher education, and student teaching in particular, as a leverage point for growing the instructional skills of teacher candidates. Importantly, this is *causal* 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, and we provide some of the first evidence. Moreover, there is a lack of quantitative evidence on how internships and mentoring should be structured to further the hands-on 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. Thus, 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.

References

- American Association of Colleges of Teacher Education (AACTE). (2018). *A pivot towards clinical practice, its lexicon, and the renewal of educator preparation: A report of the AACTE Clinical Practice Commission*. Washington, DC: Author.
- American Association of Colleges of Teacher Education (2010). *AACTE Endorses NCATE Blue Ribbon Panel Report on Clinical Preparation*. Retrieved from <https://secure.aacte.org/news-room/press-releases-statements/90-aacte-endorses-ncate-blue-ribbon-panel-report-on-clinical-preparation>
- 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.
- 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.
- 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.
- Burgess, S., Rawall, S., & Taylor, E. S. (2019). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. Working Paper). Cambridge, MA. Retrieved from <https://scholar.harvard.edu/files/erictaylor/files/teacherpeer-obsv-brt-jan-19.pdf>.
- CAEP. (n.d.). Standard 2: Clinical Partnerships and Practice. Retrieved October 10, 2019, from <http://www.caepnet.org/standards/standard-2>.

- Campbell, S. L., & Ronfeldt, M. (2018). Observational Evaluation of Teachers: Measuring More Than We Bargained for? *American Educational Research Journal*, 55(6), 1233–1267.
- 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 (No. 223-1019). *CALDER Working Paper*.
- 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.
- Connor, K., & Killmer, N. (1995). Evaluation of cooperating teacher effectiveness. Paper presented at the annual meeting of the Midwest Educational Research Association, Chicago, IL.
- Cuenca, A., Schmeichel, M., Butler, B. M., Dinkelman, T., & Nichols Jr, J. R. (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.
- Florida Administrative Code (FAC). 6A-5.065. The Educator Accomplished Practices. February 13, 2011.
- Florida Department of Education. (2019). The Florida Educator Accomplished Practices (FEAPs). Retrieved October 18, 2019, from <http://www.fldoe.org/teaching/professional-dev/the-fl-educator-accomplished-practices.stml>.
- Ganser, T. (2002, December). How teachers compare the roles of cooperating teacher and mentor. In *The educational forum* (Vol. 66, No. 4, pp. 380-385). Taylor & Francis Group.
- Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals. (NCEE 2018-*

- 4001). Washington, DC: National Center for Education Evaluation and Regional Assistance. Retrieved from <https://ies.ed.gov/ncee/pubs/20184001/pdf/20184001.pdf>.
- Gareis, C. R., & Grant, L. W. (2014). The efficacy of training cooperating teachers. *Teaching and Teacher Education*, 39, 77-88.
- Glenn, W. J. (2006). Model versus mentor: Defining the necessary qualities of the effective cooperating teacher. *Teacher education quarterly*, 33(1), 85-95.
- Goldhaber, D., Krieg, J. M., & Theobald, R. (2017). Does the match matter? Exploring whether student teaching experiences affect teacher effectiveness. *American Educational Research Journal*, 54(2), 325-359.
- Goldhaber, D., Krieg, J., Naito, N., & Theobald, R. (2019). Making the most of student teaching: The importance of mentors and scope of change. *Education Finance and Policy*, 1-21.
- Goldhaber, D., Krieg, J., & Theobald, R. (2020a). Effective like me? Does having a more productive mentor improve the productivity of mentees? *Labour Economics*, 63, 101792.
- Goldhaber, D., & Ronfeldt, M. (2020b). Toward causal evidence on effective teacher preparation. In J. E. Carinici, S. J. Meyer, & C. J. Bellwether (Eds.), *Linking Teacher Preparation Program Design and Implementation to Outcomes for Teachers and Students* (pp. 211-236).
- Graham, B. (2006). Conditions for successful field experiences: Perceptions of CTs. *Teaching and teacher education*, 22(8), 1118-1129.
- Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (Measures of Effective Teaching Project). Seattle, WA: 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*, 99-112.
- 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.
- Matsko, K.K., Ronfeldt, M., Green Nolan, H., Klugman, J., Reiningger, M., Brockman, S.L. (2018). Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education*. Advance online publication. DOI: [10.1177/0022487118791992](https://doi.org/10.1177/0022487118791992).
- National Council for Accreditation of Teacher Education (NCATE). (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). Washington, DC: NCATE.
- National Council on Teacher Quality (NCTQ). (2016). *A closer look at student teaching: Undergraduate elementary programs*. Washington, DC: NCTQ.
- National Council on Teacher Quality (NCTQ). (2017). *A closer look at student teaching: Undergraduate secondary programs*. Washington, DC: NCTQ.
- 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*, 105-119.

- 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), 359-88.
- Pritchard, A. (2017). *Ways of learning: Learning theories for the classroom*. Routledge.
- Rice, J. K. (2010). *The Impact of Teacher Experience: Examining the Evidence and Policy Implications*. Brief No. 11. National center for analysis of longitudinal data in education research.
- Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education*, 66(4), 304-320.
- Ronfeldt, M., Brockman, S., & Campbell, S. (2018a). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher*.
- Ronfeldt, M., Goldhaber, D., Cowan, J., Bardelli, E., Johnson, J., & Tien, C. D. (2018b). Identifying promising clinical placements using administrative data: Preliminary results from ISTI Placement Initiative Pilot. CALDER Working Paper No. 189.
- Ronfeldt, M., Matsko, K.K., Greene Nolan, H., & Reininger, M. (2018c). Who knows if our teachers are prepared? Three different perspectives on graduates' instructional readiness and the features of preservice preparation that predict them (CEPA Working Paper No.18-01). Retrieved from Stanford Center for Education Policy Analysis: <https://cepa.stanford.edu/wp18-01>.
- 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.

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.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American*

Economic Review, 102(7), 3628-51.

Vagi, R., Pivovarova, M., & Barnard, W. (2019). Dynamics of preservice teacher quality.

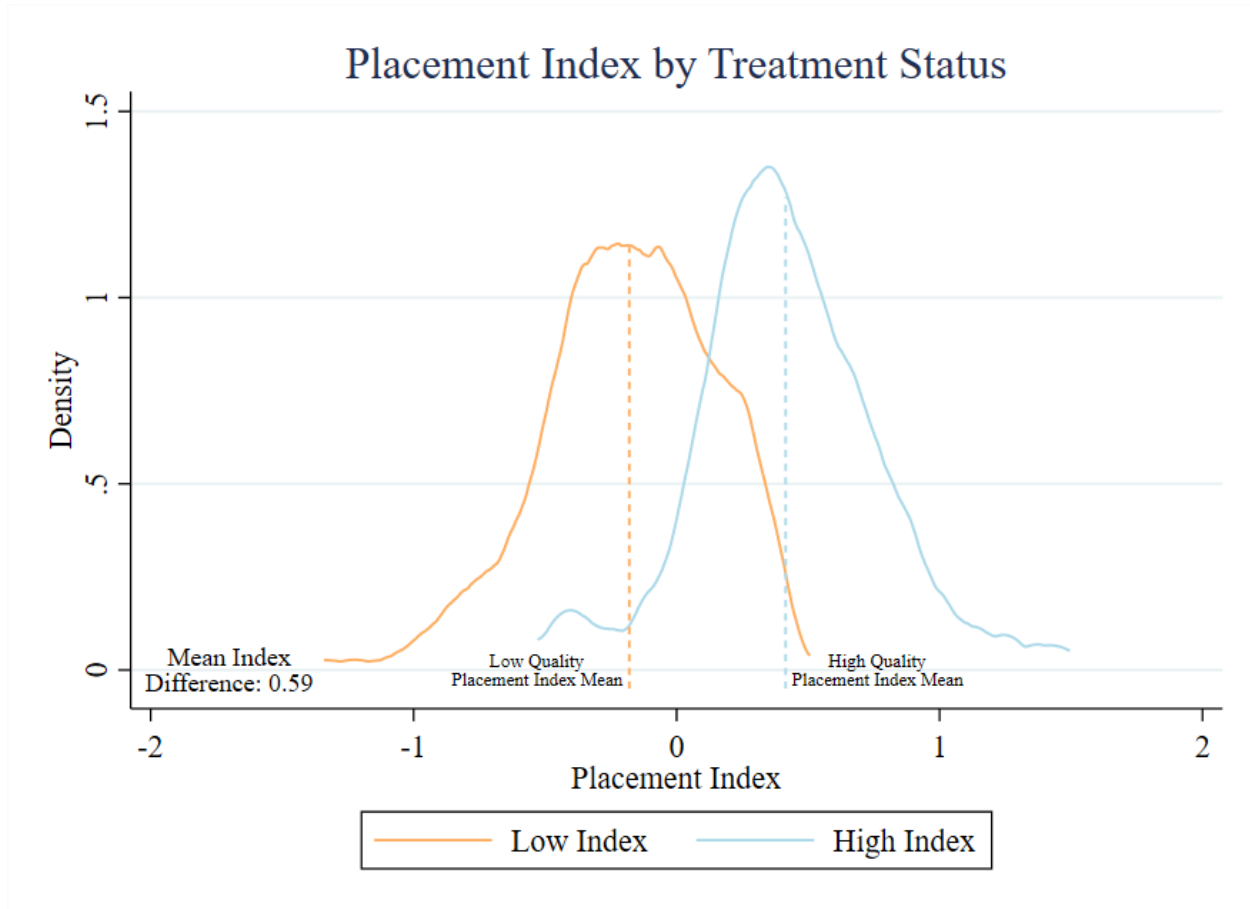
Teaching and Teacher Education, 85, 13-23.

Zeichner, K. M. (2009). *Teacher education and the struggle for social justice*. Routledge, New

York, NY.

Figures and Tables

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. High Index teacher candidates were randomized to high quality mentor teachers i.e. were treated. Randomization was done within geographic-grade-subject cells.

Figure 2. Changes in Evaluation Ratings over the Internship



Notes: 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 five-point scale, but used software permitting ratings on a decimal scale, we took the largest integer smaller than the recorded score for any non-integer scores. Evaluation rubric category names are taken from the Tennessee TEP.

Table 1. Summary Statistics and Balance Tests**Panel A: Teacher Candidate Characteristics**

	Placement Status			Difference	Effect Size
	Overall	Low Index	High Index		
Female	0.857	0.832	0.878	0.044	0.125
Male	0.14	0.161	0.125	-0.035	0.101
White	0.961	0.965	0.959	-0.006	0.033
Non-White	0.039	0.034	0.041	0.006	0.033
Cumulative GPA	3.61	3.63	3.6	-0.028	0.089
Inclusive GPA	3.435	3.411	3.46	0.044	0.125

Panel B: Mentor Teacher and School Characteristics

	Placement Status			Difference	
	Overall	Low Index	High Index		
Mentor VA	0.017 (0.953)	-0.456 (0.849)	0.41 (0.853)	0.866	***
Mentor Experience	0.009 (1.013)	-0.496 (0.709)	0.479 (1.028)	0.975	***
Mentor Observation Rating	0.374 (0.774)	0.013 (0.711)	0.708 (0.675)	0.695	***
School VA	0.005 (1.015)	-0.075 (1.083)	0.08 (0.944)	0.155	
School Stay Ratio	0.039 (0.992)	-0.031 (1.026)	0.104 (0.956)	0.135	
Student Teachers		180	189		

Panel C: Balance Tests

	Chi Square	df	p-value
<i>Teacher Candidate Characteristics from Panel A</i>			
Stratification by Randomization Block	6.48	5	0.262
<i>Field Instructor Indicators</i>			
Stratification by Randomization Block	44.76	44	0.44

Notes: With the exception of observational ratings in Tennessee 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.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Placement Initiative Experiment

		(1)	(2)	(3)	(4)
Experiment	2nd Evaluation	0.372*** (0.013)	0.371*** (0.018)	0.372*** (0.013)	0.370*** (0.018)
	3rd Evaluation	0.548*** (0.013)	0.520*** (0.018)	0.548*** (0.013)	0.519*** (0.018)
	High Index Placement	0.068 (0.045)	0.050 (0.047)	0.067 (0.044)	0.049 (0.046)
	High Index Placement * 2nd Evaluation		0.003 (0.025)		0.003 (0.025)
	High Index Placement * 3rd Evaluation		0.058** (0.026)		0.058** (0.026)
	Field Instructor Fixed Effects			X	X
	Observations	18080	18080	18080	18080
Student Teachers-Mentor Pairs	369	369	369	369	
Field Instructor Effects Variance	0.201	0.201	0.268	0.269	

Notes: Regressions of standardized evaluation scores 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 Initiative

* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 3: Lagged Field Instructor Evaluations and the Placement Initiative Experiment

	(1)	(2)	(3)	(4)
3rd Evaluation	0.067*** (0.014)	0.067*** (0.014)	0.066*** (0.014)	0.067*** (0.014)
Prior Evaluation Rating	0.298*** (0.009)	0.297*** (0.009)	0.298*** (0.009)	0.297*** (0.009)
High Index Placement	0.068* (0.040)	0.082** (0.038)	0.061 (0.041)	0.076** (0.038)
School Controls			X	X
Field Instructor Fixed Effects		X		X
Observations	11526	11526	11526	11526
Student Teachers-Mentor Pairs	340	340	340	340
Field Instructor Effects Variance	0.109	0.19	0.106	0.187

Notes: Regressions of standardized evaluation scores on order of evaluations and placement status. All evaluations (the dependent variable) were submitted by field instructors. Prior evaluation rating is the teacher candidates previous evaluation rating by their field instructor by item. All models include field instructor, candidate, and item by program random effects. Field instructor random effects are replaced with fixed effects where indicated.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Placement Initiative Experiment and Field Instructor Evaluations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	
2nd Evaluation	0.372*** (0.013)	0.371*** (0.018)	0.372*** (0.013)	0.371*** (0.018)	0.372*** (0.013)	0.371*** (0.018)	0.372*** (0.013)	0.371*** (0.018)	0.372*** (0.013)	0.371*** (0.018)	0.372*** (0.013)	0.371*** (0.018)	
3rd Evaluation	0.548*** (0.013)	0.520*** (0.018)	0.548*** (0.013)	0.520*** (0.018)	0.548*** (0.013)	0.520*** (0.018)	0.548*** (0.013)	0.520*** (0.018)	0.548*** (0.013)	0.520*** (0.018)	0.548*** (0.013)	0.520*** (0.018)	
Experiment	High Index Placement	0.068 (0.045)	0.050 (0.047)	0.084* (0.047)	0.066 (0.049)	0.011 (0.051)	-0.007 (0.053)	0.039 (0.052)	0.022 (0.053)	0.062 (0.046)	0.045 (0.048)	0.063 (0.046)	0.046 (0.047)
	High Index Placement * 2nd Evaluation		0.003 (0.025)		0.003 (0.025)		0.003 (0.025)		0.003 (0.025)		0.003 (0.025)		0.003 (0.025)
	High Index Placement * 3rd Evaluation		0.058** (0.026)		0.058** (0.026)		0.058** (0.026)		0.058** (0.026)		0.058** (0.026)		0.058** (0.026)
MT Characteristics	Mentor Value-Added			-0.053* (0.031)	-0.053* (0.031)								
	Mentor Exp. (std)					0.060** (0.027)	0.060** (0.027)						
	Mentor Standardized Observational Rating(OR)							0.039 (0.038)	0.038 (0.038)				
School	School Value-Added									-0.001 (0.026)	-0.000 (0.026)		
	School Retention											0.030 (0.032)	0.030 (0.032)
Observations	18080	18080	18080	18080	18080	18080	18080	18080	18080	18080	18080	18080	18080
Student Teachers-Mentor Pairs	369	369	369	369	369	369	369	369	369	369	369	369	369
Field Instructor Effects Variance	0.201	0.201	0.199	0.2	0.205	0.206	0.201	0.201	0.202	0.203	0.198	0.199	0.199

Notes: Regressions of standardized evaluation scores 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 Initiative

* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 5. Domain Analyses and the Placement Initiative Experiment

	(1)	(2)	
2nd Evaluation	0.372***		
	(0.013)		
3rd Evaluation	0.547***	0.078***	
	(0.013)	(0.014)	
Domain: Environment	0.411***	0.263***	
	(0.025)	(0.026)	
Domain: Planning	-0.101***	-0.105***	
	(0.026)	(0.026)	
Prior Evaluation Rating		0.260***	
		(0.010)	
Prior Evaluation Rating * Environment		0.069***	
		(0.018)	
Prior Evaluation Rating * Planning		-0.043**	
		(0.019)	
Experiment	High Index Placement * Instruction	0.080*	0.081*
		(0.046)	(0.043)
	High Index Placement * Environment	0.031	0.025
		(0.054)	(0.051)
	High Index Placement * Planning	0.053	0.074
		(0.054)	(0.051)
Observations	18080	18080	
Student Teachers-Mentor Pairs	369	340	
Field Instructor Effects Variance	0.198	0.118	

Notes: Regressions of standardized evaluation scores on order of evaluations, Placement Status, and when indicated prior evaluation rating. Order indicates the observational order of the evaluation within candidate-item-field instructor cells. The base domain is "Instruction", so main effects should be interpreted as effects for the "Instruction" domain. 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 Initiative

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Differences in Evaluation Scores by Field Instructors and Mentor Teachers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Experiment	High Index Placement	0.016			0.065			
		(0.101)			(0.100)			
Mentor Characteristics	Mentor Value-Added		-0.160**			-0.150**		
			(0.061)			(0.067)		
	Mentor Exp. (std)			0.033			0.046	
				(0.049)			(0.059)	
	Mentor Standardized Observational Rating(OR)				-0.002			0.056
					(0.066)			(0.072)
	Observations	2023	2023	2023	2023	1726	1726	1726
	Student Teachers-Mentor Pairs	76	76	76	76	66	66	66
	Observation Submitted within X days	7	7	7	7	0	0	0

Notes: Regression of the difference between mentor teacher 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 < 0.10, ** p < 0.05, *** p < 0.01.

Appendix A: Evaluation Ratings by Instructor Type

We explore the rating differences between field instructors and mentor teachers in Appendix Table A1. Unlike our main models, Equations (1)-(4), here we stack both mentor teacher and field instructor evaluation ratings. To assess these relationships, we use variations of the following equation:

$$(A1) \quad Y_{itjp} = \beta_1 \text{Mentor} + \beta_2 \text{Order}_{itjp} + \alpha_{tp} + \theta_j + \omega_i + \epsilon_{itjp}$$

In Eq. (A1), i denotes individual, l denotes the item, t denotes the ordinal position of the rating, j denotes the field instructor, and p denotes the TEP. *Mentor* is a 0/1 indicator for whether or not the evaluation came from a mentor teacher instead of a field instructor. *Order* denotes the ordinal position of the rating within candidate-mentor/field instructor cells. In Eq. (A1), we include item-by-program, student teacher, and field instructor random effects.

Appendix Table A1: Evaluation Ratings by Instructor Type

	(1)	(2)
Mentor	-0.033*	0.116***
	(0.019)	(0.019)
2nd Evaluation		0.388***
		(0.012)
3rd Evaluation		0.561***
		(0.013)
Observations	20444	20444

Notes: Regressions of standardized evaluation scores, time elapsed, and instructor type. Order indicates the order of the evaluation within candidate-item-mentor cells. Mentor indicates an evaluation by a mentor teacher. All models include field instructor, candidate, and item by program random effects.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix B: Mentor Teacher Evaluations by Mentor Characteristics

We investigate the first evaluation ratings of teacher candidates conducted by mentor teachers as they relate to mentor teacher characteristics in Appendix Equation (B2), and we display these results in Appendix Table B2. This analysis is dissimilar to our main models in that it only utilizes mentor teacher evaluation scores rather than field instructor evaluation ratings.

These relationships are explored by the following equation:

$$(B2) \quad Y_{ilp} = \beta_1 Mentor_i + \alpha_{lp} + \omega_i + \epsilon_{ilp}$$

In Eq. (1), i denotes individual, l denotes item, and p denotes the TEP. $Mentor_i$ is a measure of mentor teacher quality; effectiveness, experience, or performance rating. In Eq. (B2), we include item-by-program fixed effects and teacher candidate random effects.

Appendix Table B2: 1st Time Mentor Teacher Evaluations by MT Characteristics

	(1)	(2)	(3)
MT VA	-0.121 (0.094)		
MT Exp.		0.089 (0.076)	
MT Observational Rating			0.089 (0.101)
Observations	1373	1373	1373
MTs	78	78	78

Note: Regressions of 1st time standardized evaluation scores by mentor teacher on mentor teacher characteristics. All models include program-item fixed effects and teacher candidate random effects.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix C: Effects of Mentor Teacher and School Characteristics on Field Instructor Evaluations

Appendix Table C3 reports specifications similar to Equation (2), but with the variable of interest, *Treat*, being replaced by mentor teacher characteristics. Appendix Table C3 shows specifications without (in columns 1-6) and with (in columns 7-12) the school characteristics (school value added and teacher retention) that form 25% of the index used to determine the field placement in the experiment. In odd columns field instructor random effects have been swapped out for field instructor fixed effects.

Findings from the table are consistent with estimates found when entering mentor and school characteristics sequentially into models that control for the type of placement. While not statistically significant in this specification, teacher candidates assigned to mentor teachers with higher value added do worse on clinical evaluation ratings, all else equal. Mentor teacher experience and mentor teacher observational rating are both positively associated with higher teacher candidate clinical evaluation ratings by field instructors, with mentor teacher experience being statistically significant. These coefficient estimates are largely unaffected by exchanging field instructor random effects for fixed effects, or the inclusion of school level controls.

In Appendix Table C4, we show the coefficients for a set of analogous models, but where the variables of interest are the school level controls. The coefficient on the school retention is positive, and depending on the specification, marginally statistically significant. The coefficient on school value added, on the other hand, is close to zero in all specifications. In Appendix Table C5, we present analogous models from Appendix Tables C3 and C4, but run models only on complete cases. That is, we omit cases where placement characteristics were mean imputed.

Appendix Table C3: Lagged Field Instructor Rating: Mentor Teacher Characteristics

Lagged Field Instructor Rating: Mentor Teacher Characteristics		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	
Mentor Characteristics	3rd Evaluation	0.067*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.066*** (0.014)	0.067*** (0.014)	0.066*** (0.014)	0.067*** (0.014)	0.066*** (0.014)	0.067*** (0.014)	
	Prior Evaluation Rating	0.298*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.296*** (0.009)	
	Mentor Value-Added	-0.012 (0.026)	-0.008 (0.025)						-0.016 (0.026)	-0.011 (0.025)				
	Mentor Exp. (std)			0.034 (0.021)	0.041** (0.020)						0.033 (0.021)	0.041** (0.020)		
	Mentor Standardized Observational Rating(OR)					0.043 (0.029)	0.040 (0.028)						0.037 (0.030)	0.038 (0.029)
	School Controls								X	X	X	X	X	X
	Field Instructor Fixed Effects		X		X			X		X		X		X
	Observations	11526	11526	11526	11526	11526	11526	11526	11526	11526	11526	11526	11526	11526
	Student Teachers-Mentor Pairs	340	340	340	340	340	340	340	340	340	340	340	340	340
	Field Instructor Effects Variance	0.11	0.185	0.111	0.19	0.107	0.183	0.108	0.183	0.107	0.187	0.104	0.18	

Notes: Regressions of standardized evaluation scores on order of evaluations and mentor characteristics. All evaluations (the dependent variable) were submitted by field instructors. Prior evaluation rating is the teacher candidates previous evaluation rating by their field instructor by item. All models include field instructor, candidate, and item by program random effects. Field instructor random effects are replaced with fixed effects where indicated.

* p < 0.10, ** p < 0.05, *** p < 0.01.

Appendix Table C4: Lagged Field Instructor Rating: School Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3rd Evaluation	0.066*** (0.014)	0.067*** (0.014)	0.066*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.068*** (0.014)	0.067*** (0.014)	0.068*** (0.014)
Prior Evaluation Rating	0.298*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.297*** (0.009)	0.297*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.296*** (0.009)
School Value-Added	-0.004 (0.024)	-0.006 (0.023)			-0.013 (0.023)	-0.015 (0.023)		
School Retention			0.048* (0.028)	0.042 (0.028)			0.039 (0.028)	0.032 (0.028)
Mentor Teacher Controls					X	X	X	X
Field Instructor Fixed Effects		X		X		X		X
Observations	11526	11526	11526	11526	11526	11526	11526	11526
Student Teachers-Mentor Pairs	340	340	340	340	340	340	340	340
Field Instructor Effects Variance	0.107	0.182	0.103	0.181	0.113	0.191	0.11	0.19

Notes: Regressions of standardized evaluation scores on order of evaluations and school characteristics. All evaluations (the dependent variable) were submitted by field instructors. Prior evaluation rating is the teacher candidates previous evaluation rating by their field instructor by item. All models include field instructor, candidate, and item by program random effects. Field instructor random effects are replaced with fixed effects where indicated.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix Table C5: Complete Placement Characteristics and Lagged Field Instructor Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	
Mentor Characteristics	3rd Evaluation	0.022 (0.019)	0.023 (0.019)	0.064*** (0.014)	0.065*** (0.014)	0.065*** (0.014)	0.065*** (0.014)	0.063*** (0.015)	0.064*** (0.015)	0.061*** (0.014)	0.062*** (0.014)	0.022 (0.020)	0.023 (0.020)
	Prior Evaluation Rating	0.286*** (0.012)	0.285*** (0.012)	0.300*** (0.009)	0.299*** (0.009)	0.299*** (0.009)	0.298*** (0.009)	0.301*** (0.009)	0.299*** (0.009)	0.300*** (0.009)	0.299*** (0.009)	0.280*** (0.012)	0.280*** (0.012)
	Mentor Value-Added	0.006 (0.027)	0.003 (0.025)									-0.004 (0.031)	-0.007 (0.028)
	Mentor Exp. (std)			0.042** (0.021)	0.049** (0.020)							-0.016 (0.033)	-0.001 (0.030)
	Mentor Standardized Observational Rating(OR)					0.048 (0.030)	0.046 (0.028)					0.065 (0.046)	0.077* (0.043)
	School Value-Added							-0.005 (0.024)	-0.005 (0.023)			-0.005 (0.036)	-0.017 (0.033)
School	School Retention								0.047 (0.029)	0.040 (0.030)	0.032 (0.039)	0.009 (0.040)	
	Field Instructor Fixed Effects		X		X		X		X		X		X
Observations	6116	6116	10756	10756	10721	10721	10228	10228	10494	10494	5676	5676	
Student Teachers-Mentor Pairs	181	181	317	317	316	316	302	302	309	309	168	168	
Field Instructor Effects Variance	0.111	0.177	0.113	0.190	0.110	0.184	0.107	0.182	0.102	0.180	0.100	0.176	

Notes: Regressions of standardized evaluation scores on order of evaluations and mentor characteristics. All evaluations (the dependent variable) were submitted by field instructors. Prior evaluation rating is the teacher candidates previous evaluation rating by their field instructor by item. All models include field instructor and candidate random effects, and item by program fixed effects. Field instructor random effects are replaced with fixed effects where indicated.

* p < 0.10, ** p < 0.05, *** p < 0.01.

Appendix D: Quasi-Experimental Mentor Teacher Characteristic Models

Results regarding mentor teacher characteristics from Table 4 compare within treatment list differences in mentor teacher characteristics to differences in the clinical evaluations of teacher candidates supervised by these mentor teachers. Since teacher candidates were not explicitly randomized to different mentor teacher characteristics (recall teacher candidates were randomized within randomization blocks, i.e. grade and subject combinations), there are two more potential sources of bias that were addressed by models that only evaluate the treatment status of an individual, but are not addressed when looking at specific mentor teacher characteristics.

Bias could be introduced by non-random sorting of teacher candidates to mentor teachers. More specifically, prior to randomization, teacher candidates could sort by ability into different grade-subject cells based off the performance of the average inservice teacher within those grade-subject cells. If, for example, more motivated teacher candidates choose their preferred grade and subject based off of the attributes of inservice teacher candidates within those blocks, then a cross-sectional estimate of the effect of mentor teacher characteristics would be biased upwards. Such bias would be mitigated by controlling for randomization block effects in the regression models, and is addressed by inverse probability weighting in the models estimating intent to treat effects. Second, after teacher candidates choose a block and are randomized to a placement list, more motivated teacher candidates may advocate to be placed with higher quality inservice teachers within their placement block. This form of bias would still affect models using mentor teacher characteristics, but not the intent to treat treatment indicator, in our baseline models. If either of these scenarios occurred, estimates of the relationship between mentor

teacher characteristics and clinical evaluations would be biased upwards. We address these concerns with the following model:

$$\overline{Y_{xbltp}} = \beta_1 \overline{Index_{xbp}} + \beta_2 \overline{Order_{xbltp}} + \beta_3 \overline{Y_{xbl(t-1)p}} + \alpha_{lp} + \theta_b + \epsilon_{xbltp} \quad (D1)$$

In Eq. (D1), x denotes the treatment list (high or low), b denotes the placement block, l denotes the item, t denotes the timing of the observation (the ordinal position of the rating), and p denotes the TEP. $\overline{Y_{xbltp}}$ is the item level block-treatment cell average score on clinical evaluation t at program p . Depending on the specification, $\overline{Index_{xbp}}$ adds the variables of interest, mentor characteristics, which have been averaged across the placement block-treatment cells. $\overline{Order_{it}}$ is the ordinal position (i.e. by date) of the rating within block-treatment-item cells. $\overline{Y_{xbl(t-1)p}}$ is similar to $\overline{Y_{xbltp}}$, but is the prior average clinical evaluation score. Equation (D1) includes item-by-program (α_{lp}) and placement block fixed effects (θ_b). The error terms, ϵ_{xbltp} , have been clustered at the TEP-block-treatment cell level.

By comparing differences in mentor teacher characteristics across placement lists within the same block, placement block fixed effects (θ_b) address the first concern that teacher candidates prior to randomization sort based on attributes of inservice teachers. Second, aggregation to the treatment-block cell level by using average clinical evaluations and average mentor teacher characteristics prevents within placement list sorting bias because sorting cancels out within a randomization treatment-block. Equation (D1) identifies the effects of mentor teacher characteristics on clinical evaluations by comparing placement blocks with larger treatment/control differences in mentor characteristics to the treatment/control differences in clinical evaluations of teacher candidates.

Results from Equation (D1) are presented in Appendix table D6. Columns (1) through (3) present results without school placement controls, and columns (4) through (6) add these placement characteristics to the model. The results for mentor teacher value added and experience, in columns (1) and (2), while not statistically significant, are directionally consistent with the results presented in Table 4. On the other hand, the results for mentor observational rating in columns (3) and (6) are mixed.

Appendix Table D6: Quasi-Experimental Models of Mentor Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
3rd Evaluation	0.047 (0.041)	0.045 (0.041)	0.044 (0.041)	0.045 (0.041)	0.047 (0.041)	0.047 (0.041)
Prior Evaluation Rating	0.308*** (0.028)	0.303*** (0.027)	0.306*** (0.027)	0.295*** (0.027)	0.292*** (0.027)	0.294*** (0.027)
Mentor Value-Added	-0.055 (0.036)			-0.026 (0.029)		
Mentor Exp. (std)		0.028 (0.028)			0.026 (0.025)	
Mentor Standardized Observational Rating(OR)			-0.015 (0.039)			0.020 (0.038)
School Controls				X	X	X
Observations	4007	4007	4007	4007	4007	4007
Block-TX Cells	117	117	117	117	117	117

Mentor Characteristics

Notes: Regressions of average standardized evaluation scores on order of evaluations and average mentor characteristics. Averages are conducted at the Block-Treatment-Item-Order level for evaluations and the Block-Treatment level for mentor characteristics. All evaluations (the dependent variable) were submitted by field instructors. Prior evaluation rating is the teacher candidates previous evaluation rating by their field instructor by item. All models include block and program-item fixed effects. Errors are clustered at the TEP-Block-Treatment level.

* p < 0.10, ** p < 0.05, *** p < 0.01.

Appendix E: Clinical Evaluation Level Analyses

The analyses presented in the body of the text leverages the item level evaluations of each teacher candidate. That is, for any given clinical evaluation by a field instructor, teacher candidates are evaluated on 18 to 19 (TEP dependent) different items. In this Appendix, we present results that utilize the average item level scores within a clinical evaluation. Doing so potentially prevents problems with clustering using item-level data and might produce more reliable standard errors. This approach is depicted in Equation (E1):

$$\overline{Y_{itjp}} = \beta_1 Treat_i + \beta_2 Order_{itj} + \beta_3 Treat_i * Order_{itj} + \theta_j + \omega_i + \epsilon_{itjp} \quad (E1)$$

In some specifications we also include the lagged average field instructor's clinical evaluation score ($\overline{Y_{it(t-1)jp}}$). Equation (E1) is estimated using candidate random effects nested under field instructor random effects, and when indicated we replace field instructor random effects with fixed effects.

Appendix Table E7: Clinical Evaluation Level Analyses

	(1)	(2)	(3)	(4)	(5)	(6)
3rd Evaluation	0.550*** (0.027)	0.551*** (0.027)	0.527*** (0.038)	0.526*** (0.038)	-0.072** (0.032)	-0.046 (0.031)
Prior Evaluation Rating					0.662*** (0.029)	0.601*** (0.030)
High Index Placement	0.068 (0.045)	0.071* (0.043)	0.055 (0.054)	0.057 (0.052)	0.042 (0.031)	0.061** (0.030)
High Index Placement * 2nd Evaluation			-0.004 (0.053)	-0.002 (0.053)		
High Index Placement * 3rd Evaluation			0.046 (0.054)	0.050 (0.054)		
Field Instructor Fixed Effects		X		X		X
Observations	1028	1028	1028	1028	659	659
Student Teachers-Mentor Pairs	369	369	369	369	340	340
Field Instructor Effects Variance	0.203	0.272	0.204	0.272	0.038	0.107

Notes: Regressions of the mean of the standardized evaluation scores on order of evaluations and placement status. All evaluations (the dependent variable) were submitted by field instructors. Prior evaluation rating is the teacher candidates previous evaluation rating by their field instructor by item. All models include field instructor, candidate, and item by program random effects. Field instructor random effects are replaced with fixed effects where indicated.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix F: Placement Characteristics Interacted with the Treatment

In the Table (F8) below, we present results from the models depicted in Equation (2), but add placement characteristics to the regressions and interact these characteristics with a teacher candidate's treatment status.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
3rd Evaluation	0.067*** (0.014)	0.068*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.066*** (0.014)	0.067*** (0.014)	0.067*** (0.014)	0.067*** (0.014)
Prior Evaluation Rating	0.298*** (0.009)	0.297*** (0.009)	0.298*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.296*** (0.009)	0.298*** (0.009)	0.297*** (0.009)	0.298*** (0.009)	0.297*** (0.009)
High Index Placement	0.060 (0.042)	0.075* (0.039)	0.057 (0.049)	0.066 (0.047)	0.041 (0.053)	0.063 (0.051)	0.077* (0.043)	0.089** (0.040)	0.077* (0.042)	0.092** (0.039)
Mentor Value-Added	-0.028 (0.043)	-0.032 (0.041)								
High Index Placement *Mentor Value-Added	0.050 (0.056)	0.050 (0.053)								
Mentor Exp. (std)			-0.031 (0.044)	-0.021 (0.042)						
High Index Placement * Mentor Exp.			0.080 (0.052)	0.071 (0.050)						
Mentor Standardized Observational Rating(OR)					0.020 (0.045)	0.005 (0.043)				
High Index Placement * Mentor Observational Rating					0.014 (0.060)	0.018 (0.057)				
School Value-Added							0.019 (0.030)	0.018 (0.029)		
High Index Placement * School Value-Added							-0.059 (0.043)	-0.062 (0.040)		
School Retention									0.084** (0.035)	0.075** (0.034)
High Index Placement * School Retention									-0.086** (0.041)	-0.083** (0.039)
Field Instructor Fixed Effects		X		X		X		X		X
Observations	11526	11526	11526	11526	11526	11526	11526	11526	11526	11526
Student Teachers-Mentor Pairs	340	340	340	340	340	340	340	340	340	340
Field Instructor Effects Variance	0.109	0.189	0.110	0.190	0.109	0.190	0.110	0.186	0.106	0.185

Notes: Regressions of standardized evaluation scores on order of evaluations and mentor characteristics. All evaluations (the dependent variable) were submitted by field instructors. Prior evaluation rating is the teacher candidates previous evaluation rating by their field instructor by item. All models include field instructor and candidate random effects, and item by program fixed effects. Field instructor random effects are replaced with fixed effects where indicated.

* p < 0.10, ** p < 0.05, *** p < 0.01.

Appendix G: Treatment Models with Block Fixed Effects

In the Tables (G9) and (G10) below, we present experimental results from the models depicted in Equations (1) and (2), but add randomization block fixed effects.

Appendix Table G9

	(1)	(2)	(3)	(4)	
2nd Evaluation	0.373*** (0.013)	0.372*** (0.018)	0.373*** (0.013)	0.372*** (0.018)	
3rd Evaluation	0.549*** (0.013)	0.521*** (0.018)	0.549*** (0.013)	0.521*** (0.018)	
Experiment	High Index Placement	0.075* (0.043)	0.058 (0.045)	0.082** (0.041)	0.065 (0.043)
	High Index Placement * 2nd Evaluation		0.003 (0.025)		0.003 (0.025)
	High Index Placement * 3rd Evaluation		0.057** (0.026)		0.057** (0.026)
	Field Instructor Fixed Effects			X	X
Block Fixed Effects	X	X	X	X	
Observations	18080	18080	18080	18080	
Student Teachers-Mentor Pairs	369	369	369	369	

Notes: Regressions of standardized evaluation scores 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 Initiative.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix Table G10

	(1)	(2)	(3)	(4)
3rd Evaluation	0.067*** (0.014)	0.067*** (0.014)	0.066*** (0.014)	0.067*** (0.014)
Prior Evaluation Rating	0.298*** (0.009)	0.297*** (0.009)	0.298*** (0.009)	0.297*** (0.009)
High Index Placement	0.063* (0.038)	0.080** (0.036)	0.057 (0.038)	0.078** (0.036)
School Controls			X	X
Field Instructor Fixed Effects		X		X
Block Fixed Effects	X	X	X	X
Observations	11526	11526	11526	11526
Student Teachers-Mentor Pairs	340	340	340	340

Notes: Regressions of standardized evaluation scores on order of evaluations and placement status. All evaluations (the dependent variable) were submitted by field instructors. Prior evaluation rating is the teacher candidates previous evaluation rating by their field instructor by item. All models include field instructor, candidate, and item by program random effects. Field instructor random effects are replaced with fixed effects where indicated.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Notes

¹ In most teacher education programs, 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 teacher education programs (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 teacher education program 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 Goldhaber et al. (2020b) for a more in depth discussion of this point.

⁴ As we describe in Ronfeldt et al. (2018b), these attributes of mentors and schools have been shown to be associated with various inservice teacher outcomes.

⁵ We standardize the individual mentor teacher attributes (teacher value added, experience, and summative evaluations) to have means 0 and standard deviations of 1 and then sum them to create a teacher index. Similarly, we create a school index by summing standardized school value-added measures and school turnover measures. In each case, the value-added measures are the official measures calculated by the state for teacher evaluation or school accountability purposes (more detail on this is provided in the Data and Sample Description section below). Similarly, the observational measures are the results provided to the state as part of state-mandated teacher evaluation systems. The teacher quality measures used in the intervention are therefore similar to measures that are available in state databases in many locations. When any of the five measures are missing, we replace missing values with mean TEP-district-by-grade-by-subject block values.

⁶ 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/>, accessed 10-16-2019.

⁷ 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/>, accessed 10-18-2019.

⁸ In Florida, we use an estimate of the school average teacher experience in place of teacher retention, as retention was unavailable.

⁹ We weight our balance test by the inverse of the probability of being placed in the treatment condition and stratify by Placement block.

¹⁰ This omnibus test requires field instructors to be paired with teacher candidates randomized to both high and low quality placements. Additionally, when stratifying by randomization block this test requires field instructors to be paired with teacher candidates from at least two randomization blocks. 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 low index mentor teacher has 7.2 years of experience and the average high 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 five-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 scores 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 (FAC 6A-5.065, 2011). Source <http://www.fldoe.org/teaching/performance-evaluation/fl-state-models-of-evaluation-system.shtml>, accessed 10-18-2019.

-
- ¹⁵ 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.
- ¹⁶ They were, for instance, just 0.5%, proportion of 1st time clinical evaluation ratings.
- ¹⁷ 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.
- ¹⁸ The model includes field instructor, teacher candidate, and program-item random effects
- ¹⁹ In an alternative specification we 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 upon request.
- ²⁰ 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).
- ²¹ 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 high-list placement. Our baseline models therefore do not control for assignment block fixed effects, although results from models that do include block affects (Appendix Tables G9 and G10) are quite similar.
- ²² 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.
- ²³ Vagi et al. (2019) find that teacher candidates who enter teacher education programs with higher grade point averages tend to show greater gains on a clinical observational rubric rating during the course of their internships.
- ²⁴ A small number of observations are missing this block average in which case imputation is done at the TEP-cohort level.
- ²⁵ TEP domains are parameterized as categorical variables.
- ²⁶ The clinical ratings are estimated to increase by about 37% of a standard deviation for the 2nd observation relative to the 1st and by 55% of a standard deviation for the 3rd relative to the 1st. This is broadly consistent with the only other quantitative study that documents changes in clinical evaluation ratings during an internship: Vagi et al. (2019) find that student teachers improve (over a yearlong internship) by about 44% of a standard deviation from their first to third observation.
- ²⁷ The variance of the field instructor fixed effects, which does not account for sampling error, is about 1.7 times as large.
- ²⁸ We also run models controlling for specific mentor characteristics and simultaneously the suite of school characteristics. Results are quantitatively similar and available upon request.
- ²⁹ 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 non-linearities 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 high-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.
- ³⁰ Results are available upon request.
- ³¹ A standard deviation of teacher experience is about 8 years of experience.
- ³² These last results are available upon request.