American Educational Research Journal
Month XXXX, Vol. XX, No. X, pp. 1–46
DOI: 10.3102/0002831219872952
Article reuse guidelines: sagepub.com/journals-permissions

Will Mentoring a Student Teacher Harm My Evaluation Scores? Effects of Serving as a Cooperating Teacher on Evaluation Metrics

Matthew Ronfeldt
Emanuele Bardelli
Stacey L. Brockman
Hannah Mullman
University of Michigan

Growing evidence suggests that preservice candidates receive better coaching and are more instructionally effective when they are mentored by more instructionally effective cooperating teachers (CTs). Yet teacher education program leaders indicate it can be difficult to recruit instructionally effective teachers to serve as CTs, in part because teachers worry that serving may

---

Matthew Ronfeldt is an associate professor of educational studies at the University of Michigan School of Education, 610 East University Avenue, Ann Arbor, MI 48109; e-mail: *ronfeldt@umich.edu*. His scholarship focuses on identifying preservice and in-service factors that improve teaching quality and other teacher outcomes, particularly among teachers working with marginalized student populations, in order to inform policy and practice.

Emanuele Bardelli is a doctoral candidate in educational studies and a fellow in the causal inference in education policy research predoctoral training program at the University of Michigan School of Education. His research interests include teacher professional development, teacher learning, and instructional practices in mathematics education.

Stacey L. Brockman is a doctoral candidate in educational studies and a fellow in the causal inference in education policy research predoctoral training program at the University of Michigan School of Education. A former high school history teacher and intervention specialist, her scholarship seeks to identify educational policies and practices that support at-risk secondary students' academic and social-emotional growth. She is also interested in how teacher education can promote teaching quality and student learning.

Hannah Mullman is a doctoral student in educational studies and a fellow in the causal inference in education policy research predoctoral training program at the University of Michigan School of Education. Her research interests include pre- and in-service teacher learning, particularly as they relate to developing practices that promote justice and equity.

negatively impact district evaluation scores. Using a unique data set on over 4,500 CTs, we compare evaluation scores during years these teachers served as CTs with years they did not. In years they served as CTs, teachers had significantly better observation ratings and somewhat better achievement gains, though not always at significant levels. These results suggest that concerns over lowered evaluations should not prevent teachers from serving as CTs.

Keywords: cooperating teacher, mentor teacher, clinical preparation, teacher evaluation, teacher education

Introduction

A growing body of evidence suggests that certain characteristics of teachers' preservice training, including aspects of student teaching, are related to better workforce outcomes (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Krieg, Theobald, & Goldhaber, 2016; Ronfeldt, 2012, 2015; Ronfeldt, Schwartz, & Jacob, 2014). Of particular relevance to the present study, three new studies have found recent graduates to be more instructionally effective when they learned to teach with more instructionally effective cooperating teachers (CTs) during their preservice training (Goldhaber, Krieg, & Theobald, 2018a; Ronfeldt, Brockman, & Campbell, 2018; Ronfeldt, Matsko, Greene Nolan, & Reininger, 2018).

Yet many teacher education program (TEP) leaders and state policy makers suggest that, despite their best efforts, teacher candidates are often placed with CTs who are not instructionally effective (Greenberg, Pomerance, & Walsh, 2011). As we describe in more detail below, there are a number of possible reasons why this might be the case. At least one possible explanation, commonly cited in Tennessee, where our study takes place, is that instructionally effective teachers fear mentoring teacher candidates will negatively impact their teacher evaluations. Given substantial evidence that new teachers are far less effective than more experienced teachers, allowing a candidate to take over the classroom for part of the year may indeed affect student achievement scores. However, early empirical evidence suggests these fears may be unwarranted. Though there are no existing studies published in peer-reviewed journals, a working paper in Washington (Goldhaber, Krieg, & Theobald, 2018b) finds that there are no average effects of supervising candidates on their student achievement gains, though lower performing teachers have worse achievement gains in math.

More studies in different labor markets and policy environments are needed—like the present study in Tennessee—in order to test whether these findings are specific to the Washington context. Additionally, in Tennessee, student achievement gains are only one aspect of the teacher evaluation

system. This study also tests whether teachers' observation ratings, which receive equal weight in state evaluations, are impacted by mentoring a candidate. We also contribute to the existing empirical base by testing whether serving as a CT affects teacher evaluations in years after mentoring a candidate. We investigate this, in part, because some existing literature suggests that mentoring can function like a form of professional development for the CT (Spencer, 2007). Finally, we test whether the effects of serving as a CT are concentrated among teachers who are more or less instructionally effective or among teachers who work at specific school levels (elementary, middle, secondary).

Results from this study suggest that, compared with other years, teachers receive better observation ratings and similar achievement gains in years that they serve as CTs. We find positive effects on observation ratings for teachers across quartiles of instructional effectiveness, though effects are the most positive for teachers in the bottom quartile. When considering achievement gains, we detect small, positive effects for top-quartile teachers and small, negative (but nonsignificant) effects for bottom-quartile teachers; this is somewhat inconsistent with Goldhaber et al. (2018b), who found negative effects across quartiles and significantly negative effects in the bottom quartile. We also find the positive effects of serving as a CT on observation ratings to be concentrated among elementary teachers and the effects on student achievement gains to be similar across school level. Finally, in years after serving as a CT, teachers perform similarly on observation ratings and slightly worse on student achievement gains, though the latter results may be explained in part by student achievement gains' regression to the mean (Atteberry, Loeb, & Wyckoff, 2015).

The results of this study suggest that concerns that serving as a CT will harm teacher evaluations seem unwarranted; in fact, mentoring a candidate may increase evaluations. As TEPs and policy makers strive to recruit instructionally effective teachers to serve as CTs, this study suggests that these teachers should consider serving as CTs because, beyond benefiting the next generation of teachers, doing so may also improve their own evaluations.

Literature Review

The vast majority of existing literature on CTs focuses on the effects of CTs on teacher candidates; however, this study investigates the effects of supervising a candidate on CTs themselves. At the present moment, we know of no published articles about the latter. In order to motivate this study, we begin by focusing on growing evidence that CTs who are instructionally effective teachers significantly impact candidate learning and performance. We then review literature about who serves as a CT and how placements are made in order to illustrate why candidates are not always placed with CTs who are highly effective teachers. We conclude with

a review of a working paper and an unpublished report that are, to our knowledge, the only existing evidence for the impacts of supervising a candidate on student achievement.

The Impact of CTs on Candidates' Instructional Effectiveness

Recent evidence suggests that new teachers are more instructionally effective in their first year if, during their preservice preparation, they received mentoring from more instructionally effective CTs. In a study evaluating statewide data from Tennessee, Ronfeldt, Brockman, and Campbell (2018) found that candidates who completed their student teaching or residency in a classroom with CTs who received observation ratings of 5.0 (significantly above expectations—the highest score on Tennessee's ratings scale) performed as if they had an additional year of teaching experience when they began teaching as compared with peers whose CTs received ratings of 3.0 (at expectations). They also found the student achievement gains of candidates and their CTs to be significantly and positively correlated. Likewise, in a study using data from Chicago Public Schools, Ronfeldt, Matsko, et al. (2018) found that an increase of one point in CTs' observational ratings (on a scale of 1–4) was associated with a 0.16 point gain for their preservice candidates' ratings in their first year, an amount comparable to the average difference on observation ratings between teachers in their first year and teachers with between 2 and 5 years of experience in Chicago (Jiang & Sporte, 2016). More recently in Washington, Goldhaber et al. (2018b) also found strong, positive associations between the math student achievement gains of mentees and mentors and more modest, but still positive, associations in English Language Arts (ELA).

CT Recruitment and Selection

The literature reviewed thus far suggests that being assigned to an instructionally effective CT predicts teacher candidates becoming more instructionally effective themselves. Yet both existing qualitative literature and anecdotal evidence indicate that teacher candidates are often assigned to CTs who are not the most instructionally effective teachers in their schools or districts (Greenberg et al., 2011). There are many possible explanations for this. First, there is evidence that some TEPs privilege recruiting CTs who are known to provide good or supportive coaching to preservice candidates over recruiting the most instructionally effective teachers of P–12 students (Mullman & Ronfeldt, 2019). Additionally, recent research conducted on student teaching placements suggests that proximity to the program or the preservice candidate's home might be the most influential factor in selection of CTs, rather than instructional quality (Krieg et al., 2016; Maier & Youngs, 2009). Prior research also suggests that different stakeholders—including program staff, district leader, school administrators, and candidates

themselves—in different programs take primary responsibility for making placement decisions (Grossman, Hammerness, McDonald, & Ronfeldt, 2008; Matsko et al., 2018), and these stakeholders likely differ in terms of how much they prioritize CT instructional effectiveness as a selection criterion. Finally, district and school leaders are sometimes hesitant to select their most instructionally effective teachers to serve as CTs because this could mean rookie teachers take over instruction for their best teachers, which they fear may have negative short-term effects on student learning and achievement, especially given the rise of high-stakes testing (St. John, Goldhaber, Krieg, & Theobald, 2018). In fact, some TEP leaders indicate that principals occasionally want to put candidates in the classrooms of struggling teachers so that they can help out and serve as "an extra set of hands" (Mullman & Ronfeldt, 2019). Similarly, St. John et al. (2018) find that principals sometimes make these matches "with the hope of either supporting or motivating a [CT's] practice" (p. 14).

Most relevant to this study, though, are reports by TEP leaders, and the district and school leaders with whom they collaborate, that teachers can be hesitant to serve as CTs for concerns that their annual evaluation scores may suffer. We initially learned about these concerns anecdotally, during conversations with TDOE policy makers and TEP leaders. These concerns were subsequently confirmed during interviews—as part of a research study on the variation in clinical preparation—by TEP leaders responsible for designing and implementing clinical experiences across Tennessee (Mullman & Ronfeldt, 2019). When asked about the basis for these concerns, some mentioned an unpublished report by the SAS Institute (2014) from a pilot study in Tennessee that concluded,

> For most grades and subjects, supervising student teachers had no significant difference in terms of teacher effectiveness, particularly for teachers who are considered average or high performing. However, the initial findings do suggest that low performing teachers might have a small negative impact in their effectiveness in Mathematics and Science when supervising student teachers as compared to not supervising. This finding has potential implications for the assignment of student-teachers to licensed teachers. (p. 2)

This potential harm to evaluation scores might worry teachers of all levels of effectiveness, given the climate of high-stakes testing.

The Impact of Mentoring on CTs' Instructional Effectiveness

Despite these concerns, one might hypothesize that instructional quality would improve in classrooms with a teacher candidate/student teacher, given the higher student-to-teacher ratio, opportunities for collaborative teaching, and the introduction of potentially new knowledge/pedagogy by the teacher candidate.

We are aware, though, of only one recent working paper that has directly tested the impact of supervising a candidate on CTs' instructional performance. In Washington, Dan Goldhaber et al. (2018b) tested whether mentoring a candidate affected student achievement gains, and whether effects were heterogenous across levels of CT instructional effectiveness, as measured by teachers' value-added scores. Using data from 14 TEPs in Washington state, they found that there was no concurrent effect of mentoring a candidate on average student math or ELA achievement gains. However, there were differential effects by quartile of prior performance, namely, mentoring a candidate had a large and negative effect on students' math achievement for CTs in the lowest value-added quartile. The authors suggest that more effective CTs are able to "mitigate" the impact of letting an inexperienced candidate take over instruction in the classroom. Conversely, they also found modest, positive impacts on student math and reading performance in subsequent years of serving as a CT. In the next section, we consider more extensively different mechanisms by which mentoring a candidate might impact a teacher's concurrent and future performance.

Contributions of the Present Study

In keeping with recent calls for more replication studies in educational research (Makel & Plucker, 2014), the present study replicates and extends the Goldhaber et al. study in a different teacher labor market and state context. Like Goldhaber et al. (2018b), we are interested in the effect that mentoring a candidate has on teachers' evaluation metrics. The present study, though, also extends prior research by incorporating both value-added measures and observational ratings as our outcomes of interest. While Goldhaber and colleagues only considered value-added measures, in many states (including Tennessee) observation ratings carry equal, and sometimes greater, weight in final evaluations. Especially given prior evidence that observation ratings may be prone to rater tendencies, biases, and subjectivities (Campbell, 2014; Campbell & Ronfeldt, 2018; White, 2018), it may be that the effects of supervising a candidate on observation ratings differ from the effects on value-added measures. For example, the elevated status of being a CT may cause raters to inflate scores of teachers supervising candidates. We further extend the literature through our use of an analytic sample from Tennessee state administrative data, a state with a labor market and cultural context that differs from Washington.

Similar to Goldhaber et al. (2018b), we consider heterogeneity of effects by prior performance level as well as heterogeneity by school level. We add the latter focus because elementary teachers typically have self-contained classrooms and teach all subjects to the same group of students, whereas secondary teachers typically work with different students (classes/preps) across the day and usually specialize in terms of subject matter. These

different arrangements require different approaches and decisions about how to integrate candidates into classrooms and lead teaching responsibilities, and thus, may have different implications for impacts on a CT's own performance. Below, we elaborate on different school-level considerations regarding student teaching arrangements and potential mechanisms by which these arrangements may impact a CT's performance.

We also consider the impacts of mentoring a candidate on the CT's own learning and professional growth and interrogate whether serving as a CT changes future performance. The literature suggests that effective professional development include long-term, active learning (Desimone, 2009), and it is possible that mentoring a novice teacher meets these requirements. In fact, in a recent survey of CTs in Chicago, almost one fifth of CTs indicated that their primary reason for serving as a CT was because it helped them to improve as a teacher (Matsko, Ronfeldt, & Greene Nolan, 2019). This is further supported by a review of mentoring programs for novice teachers in the United Kingdom, where Shanks (2017) finds that mentors, in coaching novices, sometimes engage in the same kinds of critical inquiry and reflection as mentees, creating opportunities for learning for both parties. Thus, we test for lagged effects of serving as a CT on teachers' instructional performance in years after they mentored candidates.

Logic Model

While most existing research has focused on the effects of CTs on the performance of those candidates working with them, this study investigates the effects on the performance of CTs themselves. The perception among some teachers, teacher educators, and policy makers in Tennessee and elsewhere—a central motivation for this study—is that serving as a CT can harm teachers' evaluation scores[1] in the year that they serve. How might this occur? To our knowledge, there is no existing research on how serving as a CT might impact one's own evaluation scores; thus, we can only speculate. In this section, we begin by considering a number of possible mechanisms by which mentoring a candidate might impact, positively and negatively, teachers' evaluation scores in the year that they serve; after, we consider how serving as a CT might affect their future performance, during postservice years.

Our logic model (see Figure 1) begins with an assumption that observation ratings and value-added measures reflect the quality of underlying teaching skills/competencies, an assumption that is supported by a number of studies demonstrating their validity and reliability (Cantrell & Kane, 2013; Gitomer & Bell, 2013; Hill, Kapitula, & Umland, 2011; Kane & Staiger, 2012). We also acknowledge, though, that these evaluation measures are unlikely to capture all dimensions of teaching quality and are known to measure other aspects of classrooms beyond teaching quality, so are prone to

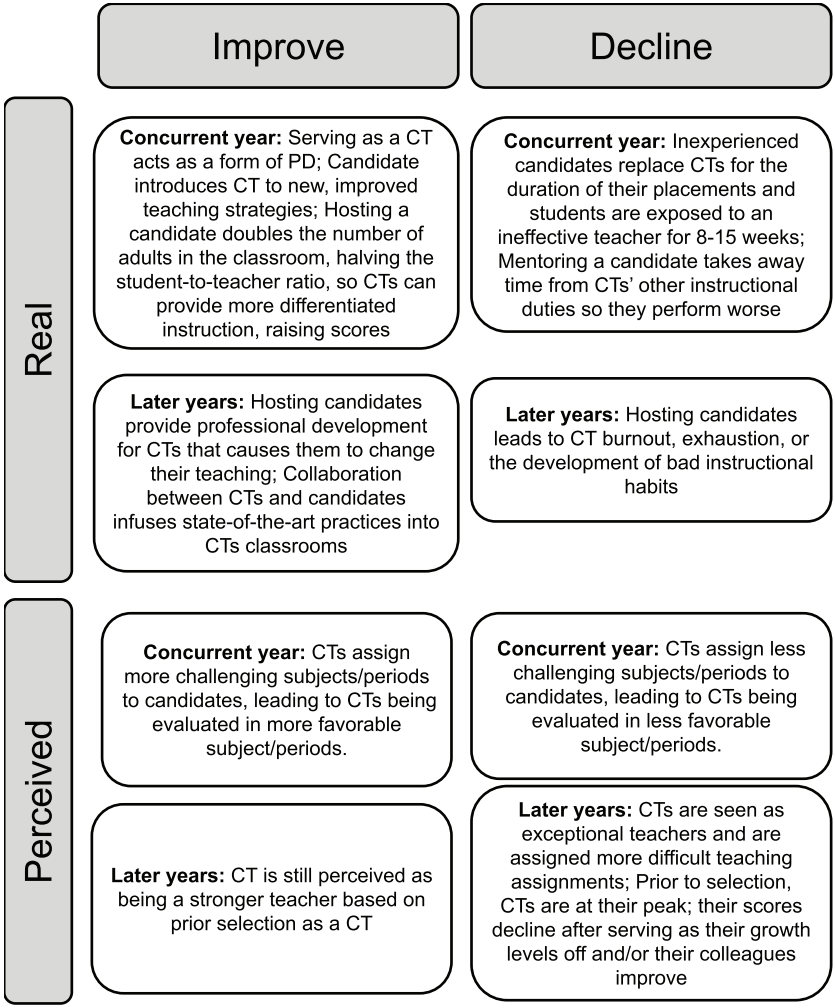

Figure 1. **Logic model for effects of serving as a cooperating teacher on evaluation scores.**

Note. CT = cooperating teacher; PD = professional development.

manipulation, error, and bias, as previously stated. We know, for example, that observation ratings tend to be lower in classrooms with students who are lower achieving, Black and Latinx, receiving special education services, and secondary, and in classrooms of teachers who are male and, in some cases, Black (Campbell & Ronfeldt, 2018; Harris, Ingle, & Rutledge, 2014;

Jiang & Sporte, 2016; Steinberg & Garrett, 2016). We also know that observation ratings may vary by time of day and year, identity of the rater (e.g., master rater versus principal), and content being taught (White, 2018; Whitehurst, Chingos, & Lindquist, 2014). Similarly, value-added measures (VAMs) are somewhat unstable across years and can vary by student characteristics, subject area, and prior achievement (Loeb & Candelaria, 2012). Thus, as we consider how ratings change in response to serving as a CT, we consider not only how serving as a CT may influence the quality of teaching in a classroom but also how evaluations may change as a result of manipulation or bias, even where the quality of teaching remains constant.

As we consider mechanisms by which CTs' performance may be affected by mentoring a candidate, it is also important to consider different ways that CTs might hand over lead teaching responsibilities to candidates. In some cases, CTs only have one classroom or section (e.g., elementary teachers in self-contained classrooms). In these cases, CTs can allow candidates to take over lead teaching responsibilities across subject areas; they can also hand over lead teaching responsibilities in some subjects (e.g., reading, science) but not others. In other cases, CTs teach multiple periods or classes (e.g., secondary science teachers with Biology and AP Biology classes). Here, CTs may hand over lead teaching responsibilities in some classes but retain them in others. Given that school principals/leaders are likely to observe CTs when they are personally teaching, the decision about which classes to hand over to a candidate can have implications for CTs' evaluations. For example, if a CT hands over lead teaching responsibilities in their most challenging classes or subjects, then they are likely to be observed and evaluated in less challenging contexts which could boost their evaluations. Thus, as we discuss mechanisms by which mentoring may impact a teacher's own evaluations, we consider ways in which the quality of teaching may be impacted versus ways that evaluated performance may change without necessarily impacting underlying teaching quality; in Figure 1 we differentiate mechanisms by which "real" and "perceived" performance may be impacted. Here we intend to differentiate ways in which teachers may actually alter their teaching (quality) in response to service as a CT versus ways in which evaluations of their practice might be altered without necessarily changing their teaching at all. Our study is not designed to test which, if any, of these postulated mechanisms is at work; rather, we include this section to give the reader an orientation to plausible ways that serving as a CT might affect a teacher's evaluations. We encourage future research to investigate which of these mechanisms explain the results we observe.

Possible Mechanisms for Impacting Performance During Service Year

We begin by considering mechanisms by which CTs' performance may be harmed in years that they mentor candidates. Perhaps the most obvious is

that candidates are inexperienced, and there is a great deal of literature demonstrating that rookie teachers tend to be less effective teachers, which could lead to lower student achievement scores. That is, when candidates take over some/all lead teaching responsibilities, students in their classrooms likely encounter less effective teaching, on average. If, as a result, students perform worse on state tests, this would be reflected in CTs' achievement gains, given that they are still the teacher of record. This is an example of a real change in teaching quality resulting from serving as a CT.

How could mentoring a candidate also negatively affect CTs' observation ratings? In Tennessee, a teacher must still be observed by a principal or school leader even when mentoring a candidate. In cases where CTs teach a self-contained class, the evaluator presumably observes the CT teaching the same students as those taught by their candidate. As a result of inexperienced candidates taking responsibility for some of the prior classroom activities, it is possible that the classroom culture will be worse and/or the students will be less prepared and, as a result, the CTs may struggle more when being observed and evaluated. In cases where CTs have multiple preps/classrooms and hand over lead teaching responsibility in only one/some classes, then it is possible that candidates will take over in classes with stronger classroom cultures (e.g., to make learning to teach somewhat easier). This would then leave CTs to be evaluated in classrooms/preps where they may be more likely to struggle and, hence, receive lower observation ratings. This scenario describes a case in which perceived teaching quality changes as a result of CT service. Yet another possibility is that mentoring a candidate takes CTs' time and effort away from improving their own teaching with P–12 students.

On the other hand, there are potential mechanisms that could lead to an improvement in CT performance. One possible mechanism is that mentoring a candidate effectively doubles the numbers of teachers in a classroom, thus decreasing the student-to-teacher ratio and raising the likelihood that students will receive more individualized attention. Moreover, mentoring a candidate allows for teacher collaboration, which can increase teaching quality and student performance; for example, candidates might share new curriculum or pedagogy perhaps from their TEPs. It is also possible that when mentoring a candidate, CTs may be more motivated to model exceptional teaching, thus putting extra time and effort into teaching. Finally, if they choose to place their candidates in their most challenging preps/classes/subjects, CTs will be more likely to be observed in settings that are favorable to their performance.

Possible Mechanisms for Impacting Performance in Future Years

It is also possible that a teacher's performance after serving as a CT may be affected. Hosting a candidate could be a form of professional development for CTs (Spencer, 2007). By observing and providing feedback to a new teacher, or in engaging in planning and reflective conferences with

them, CTs might sharpen their practice and become more effective teachers themselves. Student teachers also may bring innovative teaching practices from their methods courses that they implement during their clinical placements. CTs could learn from these practices and add them to their teaching repertoire, which could manifest in improved performance in subsequent years.

Of course, a CT's performance could also decline in postservice years. Hosting a candidate could be taxing for a CT, either personally or professionally. This could lead to CT burnout and exhaustion. Having to fulfill the always-increasing demands of teaching on top of mentoring a new teacher could lead to CTs not having enough time to properly rest, which could result in a decline in performance in the years following mentoring a candidate. It is also possible that teachers develop poor habits when mentoring a candidate and then carry these habits into future years (e.g., adopt ineffective practices used by the candidate).

School leaders might also believe that teachers who serve as CTs are exceptional, and, in subsequent years, assign these teachers more difficult preps/classes, leading to lower evaluations in subsequent years. It is also possible that school leaders inflate evaluations of teachers in years they serve as CTs (e.g., leaders may be more lenient given that being a CT is a form of service that effectively increases workload); consequently, postservice evaluations might subsequently decline mechanically even where the quality of teaching performance is consistent.

Research Questions

Drawing upon this logic model, in this article we ask broadly, "What effects might serving as a CT have on teachers' concurrent and subsequent performance?" More specifically, the following research questions guided our analysis:

Research Question 1: Do teachers perform differently in years that they serve as CTs?

Research Question 2: Are the effects different for different groups of CTs?

Research Question 3: Do teachers perform differently in years after they serve as CTs?

Data

Data for this article come from a unique data set of CTs collected by the Tennessee Department of Education. This data set includes information from 17 TEPs[2] in the state and identifies the teachers who served as CTs for these programs between the 2010–2011 and the 2013–2014 school years. We merge these data onto Tennessee's teacher and school databases. The

teacher database includes information about teachers' work experience, licensing status, and evaluation scores. School-level data come from Tennessee school universe files which include information about student body characteristics, average attendance, and school improvement status.

Descriptive Statistics

Our analytic data set includes all teachers in Tennessee from the 2010–2011 school year through the 2016–2017 school year. Our sample includes 458,717 teacher-by-year observations. Table 1 presents descriptive information about types of evaluation data we have for teachers, including observation ratings and their value-added to student achievement measures (Teacher Value-Added Assessment System, or TVAAS; see Vosters, Guranio, and Wooldridge, 2018, for more information on how these scores are calculated). Similar to other states' use of value-added measures, TVAAS is calculated using state test data and intends to capture an individual teacher's effect on student achievement; teachers receive scores for specific tested subjects as well as composite scores. TVAAS was piloted in the 2010–2011 school year and fully implemented the following year, so we report value-added measures starting in 2011. TVAAS scores are available only for about half of the teachers in our sample because of variation in testing requirements across grade levels and school settings. Observation ratings are available starting from the 2011–2012 school year. We have a total of 4,522 teacher-by-year observations for teachers who served as CTs between the 2010–2011 and 2013–2014 school years.[3] Teachers in Tennessee are assessed multiple times per year using the Tennessee educator acceleration model, a rubric that includes four domains and multiple indicators within each domain.[4] The four domains are instruction, environment, planning, and professionalism. Professionalism is only assessed one time, at the end of the school year. For the other three, multiple domains and indicators are scored simultaneously, during the same observation, and teachers receive scores on a scale from 1 (*significantly below expectations*) to 5 (*significantly above expectations*). For this article, domain and overall ratings are an average of indicator scores and domain scores, respectively.

Table 2 presents summary statistics comparing those teachers who served as CTs with those who did not. Reading the table from left to right, we present the average statistics for our entire analytic sample of teachers, CTs, all other teachers, and the difference between CTs and other teachers. Teachers who served as CTs are, on average, statistically different from other teachers when it comes to their observation ratings, TVAAS scores, teacher covariates, and school covariates. On average, we find that CTs are more likely to be White (7.6 percentage point difference) and female (3.6 percentage point difference), have 1.88 years more experience, are more likely to hold an advanced degree, and work in schools with a greater proportion

Table 1
Number of Teachers With Valid Evaluation Data

	2011	2012	2013	2014	2015	2016	2017	Total
All teachers	21,708	74,512	71,756	73,906	71,966	72,807	72,062	458,717
Observation ratings	0	70,616	65,894	57,637	60,873	70,523	69,681	395,224
TVAAS scores	21,291	21,843	30,551	31,714	25,523	8,879	21,158	160,959
Did serve as cooperating teacher	417	1,163	1,561	1,381				4,522
Observation ratings	0	1,151	1,507	1,291				3,949
TVAAS scores	417	472	983	786	*Cooperating teacher data unavailable*			2,658
Did not serve as cooperating teacher	21,291	73,349	70,195	72,525				237,360
Observation ratings	0	69,465	64,387	56,346				190,198
TVAAS scores	20,874	21,371	29,568	30,928				102,741

Note. TVAAS = Teacher Value-Added Assessment System. The Cooperating Teacher database is available for a subset of teacher education programs in the state for school years 2010–2011 through 2013–2014. The Tennessee Teacher Value-Added Assessment System was piloted during the 2010–2011 school year and fully implemented in the 2011–2012 school year. Observation ratings are available starting from the 2011–2012 school year.

Table 2
Comparing Observable Characteristics of All Teachers With Those of Cooperating Teachers

	All Teachers	Cooperating Teachers	Other Teachers	Difference	*p* Value
Outcomes of interest					
Observation ratings	3.888	4.042	3.879	0.163	***
TVAAS—All subjects	0.042	0.098	0.038	0.061	***
TVAAS—Mathematics	0.083	0.158	0.078	0.081	***
TVAAS—ELA	0.022	0.054	0.019	0.035	***
Teacher covariates					
Percent female	0.799	0.833	0.797	0.036	***
Percent White	0.870	0.941	0.866	0.076	***
Percent Black	0.122	0.055	0.125	−0.070	***
Percent other	0.005	0.004	0.005	−0.002	***
Percent bachelor's degree	0.408	0.326	0.414	−0.088	***
Percent master's degree	0.503	0.547	0.500	0.047	***
Percent PhD	0.009	0.012	0.009	0.002	*
Age	42.55	42.95	42.53	0.42	***
Years of teaching experience	11.96	13.74	11.86	1.88	***
School assignment					
Elementary school	0.433	0.498	0.429	0.068	***
Middle school	0.185	0.193	0.185	0.009	**
High school	0.278	0.230	0.280	−0.051	***
School covariates					
Percent White	0.678	0.759	0.673	0.086	***
Percent Black	0.216	0.140	0.221	−0.081	***
Percent Hispanic	0.075	0.074	0.075	−0.001	
Percent FRPL	0.587	0.567	0.589	−0.022	***
Percent proficient	0.514	0.537	0.513	0.024	***
N	241,882	4,522	237,360		

Note. ELA = English Language Arts; FRPL = free or reduced-priced lunch; TVAAS = Teacher Value-Added Assessment System. The Cooperating Teacher database is available for a subset of teacher education programs in the state for school years 2010–2011 through 2013–2014. The Tennessee Teacher Value-Added Assessment System was piloted during the 2010–2011 school year and fully implemented in the 2011–2012 school year. Observation ratings are available starting from the 2011–2012 school year.
$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

of students who are White and meet proficiency levels on state exams and with a smaller proportion of students who qualify for free or reduced-priced lunch. CTs also tend to have higher observation ratings and TVAAS scores. In our sample, the average observation rating for a CT was 4.04, compared with 3.88 for teachers who did not serve. The average TVAAS score for CTs was

0.061 student standard deviation units higher than other teachers. These findings are consistent with other prior research which has found CTs to have stronger evaluation scores, on average, than non-CTs (Goldhaber et al., 2018b; Matsko et al., 2018; Ronfeldt, Brockman, & Campbell, 2018).

CT Blocks

In order to conduct a more appropriate comparison of those who serve as CTs with those who do not, we construct blocks of all eligible teachers for a student teaching placement in a given year. We identify teachers who served each year and then group them with all other teachers in their districts with the same teaching endorsement (e.g., secondary math, elementary, secondary ELA, etc.). This allows us to create a hypothetical pool of all teachers who could have potentially served as CT for a particular candidate.[5] We merge Tennessee's Personnel Information Reporting System and teacher assignment data onto our analytic sample and then compare the courses they taught that year and assigned them an endorsement. If for example, a seventh-grade social studies teacher in district *D* serves as a CT, we create a block with all secondary social studies teachers in district *D* in order to build a sample of all possible CTs for that year.

Table 3 presents descriptive differences between blocks with and without eligible CTs, as well as differences between CTs and the rest of the teachers in their block. Compared with blocks without CTs, on average, blocks with CTs have lower observation ratings and years of experience but higher TVAAS scores. Blocks with CTs have a higher share of elementary and middle school teachers and lower share of secondary teachers. They also tend to have more female and Black teachers but fewer White teachers.

When we look within blocks, we find that CTs outperform non-CTs. In our sample, CTs have average observation ratings of 4.03, approximately 0.19 higher than non-CTs in the same blocks. CTs also have higher TVAAS scores than non-CTs (a 0.08 standard deviation difference for teachers not in blocks and a difference of 0.06 for those in blocks). CTs were more likely to be female, White, and hold a graduate degree, but were less likely to be Black. When compared with the rest of their block, CTs were also more likely, on average, to teach in schools with higher proportions of White and higher achieving students.

Method

Research Question 1

We use a generalized differences-in-differences method with teacher fixed-effects model to investigate the effects of serving as a CT on evaluation metrics. This modeling strategy allows us the estimate the within-teacher changes in years during which they serve as a CT as compared with the other

Table 3
Descriptive Statistics by Block

		Blocks With CTs		
	Blocks Without CTs	All	Non-CTs	CTs
Outcomes of interest				
Observation ratings	3.911	3.855	3.838	4.026
TVAAS	0.033	0.051	0.045	0.109
TVAAS—Mathematics	0.063	0.100	0.093	0.171
TVAAS—ELA	0.019	0.024	0.020	0.058
Teacher covariates				
Percent female	0.781	0.824	0.822	0.845
Percent White	0.878	0.857	0.849	0.940
Percent Black	0.110	0.137	0.145	0.056
Percent other	0.005	0.005	0.005	0.004
Percent bachelor's degree	0.401	0.419	0.429	0.330
Percent master's degree	0.509	0.493	0.487	0.548
Percent PhD	0.010	0.008	0.008	0.011
Age	42.87	42.10	42.05	42.69
Years of teaching experience	12.13	11.72	11.57	13.41
School assignment				
Elementary school	0.39	0.50	0.50	0.53
Middle school	0.17	0.21	0.21	0.19
High school	0.33	0.20	0.20	0.21
School covariates				
Percent White	0.693	0.657	0.648	0.752
Percent Black	0.207	0.230	0.238	0.145
Percent Hispanic	0.072	0.079	0.080	0.074
Percent FRPL	0.588	0.586	0.589	0.564
Percent proficient or above	0.517	0.509	0.507	0.537
N	102,560	118,562	114,037	4,222

Note. CT = cooperating teacher; ELA = English Language Arts; FRPL = free or reduced-priced lunch; TVAAS = Teacher Value-Added Assessment System. Blocks were calculated according to whether a CT served in particular district in a given year. We group them with all other teachers in that district with the same teaching endorsement, so blocks represent all eligible CTs for a preservice teacher that year.

years during which they did not mentor a teacher candidate while accounting for differences between teachers who are selected and teachers who are not selected to serve as a CT.

Our preferred model is

$$Y_{it}=\beta_{0i}+\beta_1 CT_{it}+\delta Exp_{it}+\lambda_t+\varepsilon_{it}, \quad (1)$$

where Y_{it} is the outcome of interest. β_{0i} is the individual-level fixed effect. CT_{it} is an indicator variable taking the value of 1 for all years t during which teacher i is reported as serving as a CT. Exp_{it} is a set of indicators for years of work experience that we add to the model to increase efficiency and account for the timing of being selected to be a CT. λ_t is the year fixed effect. We use these fixed effects to account for any secular variation in evaluation scores. ε_{it} is the stochastic error term adjusted for clustering of teachers at the school level.

Our coefficient of interest is β_1. This term captures the causal effect of serving as a CT on evaluation scores and teacher value-added estimates. Our causal claim rests on two identifying assumptions. First, any individual-level characteristics that lead to selection into serving as a CT are constant over time and can be accounted for by an individual-level fixed effect. Second, these characteristics have a linear and additive functional form to the model's intercept (Angrist & Pischke, 2008).

Research Question 2

Heterogeneity by Quartile

Goldhaber et al. (2018b) suggested that the effects of serving as a CT vary for teachers in different effectiveness quartiles, with a negative effect among the lowest quartile of teachers.[6] We calculate effectiveness quartiles using a two-step approach. First, we estimate the teacher fixed effect from this model:

$$Y_{it} = \tau_i + \pi_1 CT_{it} + \varepsilon_i, \tag{2}$$

where τ_i is the teacher fixed effect for teacher i. This captures the evaluation score averages for teacher i over all observation years, controlling for effects of serving as a CT on evaluation scores. We use these teacher fixed effects to calculate the quartile of effectiveness for each teacher or, more formally, $Q_i|\tau_i$. We use these quartiles to estimate the effect of serving as a CT for teachers across the quality distribution using the model

$$Y_{it} = \beta_{0i} + \beta_1 CT_{it} \times \mathrm{Q_i} + \delta Exp_{it} + \lambda_t + \varepsilon_{it}, \tag{3}$$

where $CT_i \times Q_i$ is the interaction term between the CT indicator and the quartile of effectiveness for each Y. β_1 is a vector of four estimates, one for each quartile of effectiveness, that allow us to test whether the effects of serving as a CT vary at different points of the teacher performance continuum.

Heterogeneity by School Level

A major difference between elementary and secondary teachers is that the former are typically with the same group of students throughout the day while the latter tend to work with different students during the school

day. As we discuss in our logic model, differences in school level could have differential impacts on CTs' evaluation scores. It is, therefore, important to investigate whether or not the effects of serving as a CT vary by school level.

Thus, we divide schools into four categories, elementary (Grades K–5), middle (Grades 6–8), high (Grades 9–12), and other (e.g., K–8 schools). We estimate a model similar to Equation (4) where we interact the CT indicator with indicators for school level, allowing us to test whether the effects of serving as a CT vary by instructional setting.

Research Question 3

We modify our preferred model described in Equation (1) to answer the third research question:

$$Y_{it} = \beta_{0i} + \beta_1 CT_{it} + \beta_2 CT_{after_{it}} + \delta Exp_{it} + \lambda_t + \varepsilon_{it}, \quad (4)$$

where we divide the counterfactual for serving as a CT in Equation (1) into two parts using the $CT_{after_{it}}$ indicator. This indicator takes the value of 1 for all teachers who were reported as being a CT for at least 1 year and for all years following serving as a CT. This allows us to separately estimate the effects of serving as a CT on evaluation metrics for the years during which a teacher serves as a CT and for the years following serving as a CT.[7] The coefficient of interest for these analyses is β_2. This captures the effects of serving as a CT in the period following this experience as compared with evaluation scores during the period preceding serving as a CT.

Robustness Checks

We run several robustness checks to test whether our results are sensitive to our model specification, to the sample of teachers that we use, and to our estimation strategy. We find that the results from our preferred model are robust against all of these checks.

First, we test whether our results are sensitive to the inclusion of teacher experience. Papay and Kraft (2015) argued that experience coefficients could be biased when used in a fixed-effects model that includes year terms. We address this concern by estimating our preferred model without the experience terms and by adjusting the experience coefficients using the technique described by Papay and Kraft (2015).

Second, we include school-level covariates to control for possible unobserved differences among workplaces that could confound selection to be a CT and evaluation scores. For example, researchers have found a relationship between teachers' evaluation ratings and the characteristics of their students (Campbell & Ronfeldt, 2018; Jiang & Sporte, 2016; Steinberg & Garrett, 2016).

Third, we estimate our preferred model on progressively more restrictive samples of teachers in order to account for various forms of likely

selection. We restrict our sample to teachers who teach in the same school district and subject area as the teachers who we observe serving as a CT, to teachers who teach in the same school and subject, and to teachers who are reported as being CTs at least once.

Last, we use traditional difference-in-differences and matched-sample model specifications to check whether our results are sensitive to model specification. In this specification, we use the equation:

$$Y_{it}=\beta_{0i}+\beta_1 CT_{ever\,it}+\beta_2 CT_{it}+\delta Exp_{it}+\pi_{\mathrm{it}}+\lambda_t+\varepsilon_{it},$$

where CT_{ever} is an indicator variable that takes the value of 1 for any teacher who served as a CT at least once, CT is the indicator variable taking the value of 1 during all years t for which teacher i is reported as serving as a CT, π_{it} is a school fixed effect, and λ_t is a year fixed-effect term. Conceptually, this difference-in-differences model compares teachers who serve as CTs with teachers who did not serve within the same school. The first difference is between teachers who ever serve as CT ("ever CTs") and teachers who never serve as a CT. This difference accounts for average unobserved differences on evaluation scores between the group of teachers that is ever selected to serve as CTs and those who were never chosen to serve. The second difference is within the group of "ever CTs" and compares the evaluation scores for the years during which these teachers serve as CT and years during which they do not. This second difference estimates the effect of serving as a CT on evaluation scores.

This difference-in-differences specification relies on more permissive assumptions than our preferred model: that the evaluation scores of teachers who were ever selected to be CTs and the ones for teachers who were not have parallel trends before CT selection. Evaluation and CT data availability make it difficult to formally assess the degree to which evaluation scores for CTs and non-CTs followed parallel trends before serving as CT. In particular, when teachers served as CTs early in our observation window, we sometimes have no data on pretrends or only scores for a single year. Thus, we limit our analysis to the 2013 and 2014 CT cohorts where we have at least 2 years of pretrend data, and we present an event study using this cohort (see Appendix Figure A1). While we observe parallel trends between CTs and non-CTs in terms of TVAAS, we observe that observation ratings (OR) seem to increase the year prior to being selected to serve as a CT.

These results could hint at CTs being selected to serve based on prior year observation score data. Given the mixed evidence during the pre-CT period and the data limitations (especially the shifting cohorts of teachers across years), it is difficult to be certain that the parallel trends assumption has been met; therefore, we only include these difference-in-differences results as a robustness check to our preferred model. We recommend caution when interpreting the results from these difference-in-differences

models, as relaxing the assumptions of our preferred model could introduce bias when the parallel trend assumptions are not met. In fact, we find that the difference-in-differences estimates have greater magnitude than our fixed-effects estimates. Two possible sources of bias can explain these results. First, if the difference-in-differences models do not meet the parallel trend assumption, the estimates will be biased upward. Second, unobserved variables could lead to an increase in observation ratings that is unrelated to serving as a CT. For example, a teacher might take on a leadership role at the school at the same time as mentoring a candidate. We could expect that taking on that leadership role could lead to higher observation ratings and that this increase is unrelated to serving as a CT.

In part because of possible concerns that, prior to serving, teachers who become CTs may be increasing on observation ratings at relatively greater rates than other teachers, we restrict the estimate sample to a matched sample of teachers. This matched-sample model allows us to construct a comparison group that is similar on observed characteristics, including pretrends on evaluation data, to teachers who serve as a CT. We do this in a two-step process. First, we identify a sample of teachers who have observed characteristics similar to our CT sample. Second, we use this matched group to calculate the effect of serving as a CT on evaluation scores. Specifically, we match CTs and non-CTs using a nearest neighbor matching algorithm that uses an exact match on teacher demographic characteristics (i.e., race/ethnicity and gender), highest level of education completed (i.e., bachelor's, postbachelor's, or master's degree), school level (i.e., elementary, middle, or high school), and CT block. We fuzzy match using Mahalanobis distance on up to two prior years of evaluation data and years of experience at the time of serving as a CT. We remove two CTs from these analyses who did not match with other teachers in the state on background characteristics. Appendix Figure A2 reports the density distributions for the fuzzy matched variables pre- and postmatching. We note that the matching procedure was able to identify similar teachers across the demographic variables for all four outcomes of interest, suggesting that these models appear to meet the common support assumption. We also note that we were not able to have quality matches on TVAAS mathematics scores 2 years prior to serving as a CT, which could introduce some bias in the matched estimates for this particular measure.

This matched-sample specification relies on the assumption that we match teachers who are reported as being CTs to teachers similar to them in all observed characteristics included in the model except for being selected to be a CT. However, these estimates could be biased if selection to be a CT is driven by unobserved teacher characteristics.

The estimates from these two alternative model specifications have generally the same sign and are larger in magnitude than the estimates from our preferred model. The results confirm that our preferred model provides the most conservative estimates of our outcomes of interest.

Results

Research Question 1: Do Teachers Perform Differently in Years That They Serve as CTs?

We present the main results for the four outcomes of interest—observation ratings, average TVAAS, mathematics TVAAS, and ELA TVAAS—in Table 4. We begin by summarizing results from our preferred models with teacher fixed effects. Across the first row, we notice that the effects of serving as a CT on evaluation metrics is either small and positive, in the case of observation ratings, or not significantly different from zero, in case of all three TVAAS estimates. Regarding observation ratings, estimates suggest that teachers' observation scores increase by 0.04 points in years that they serve as CTs as compared with other years; this is roughly equivalent to about one fifth of the expected growth in observation ratings for a first-year teacher (Ronfeldt, Brockman, & Campbell, 2018). It is worth noting that CTs have, on average, almost 14 years of experience, a point in teachers' careers when their observation ratings tend not to increase substantially (i.e., after the 10-year mark, see Papay & Kraft, 2015, for an in-depth analysis).

The even-numbered columns in Table 4 display the estimates from the difference-in-differences models. We note that the point estimates for our coefficient of interest tend to be greater in magnitude in these models than in the teacher fixed-effects ones.[8] In fact, the estimate on models for TVAAS (all subjects), is now positive and statistically significant at the 5% level, suggesting that teachers have greater achievement gains in years that they serve as CTs. These models also allow us to estimate the difference in evaluation scores (across years) between teachers who are reported as serving as CTs at least once in our data set (see row "Ever cooperating teacher") and teachers who are not reported as serving as CTs during our observation period. We interpret this coefficient as the baseline difference in evaluation scores that might have led specific teachers to be selected as CTs. Across all four outcomes, we note that teachers who serve as CTs at least once have significantly and meaningfully higher evaluation scores than their peers. In other words, teachers who serve as CTs are, on average, higher performing teachers and seem to be positively selected on their evaluation scores.

Sensitivity to Sample Selection

As discussed in the "Data" section (see Table 2), we find that teachers in the same districts and subject areas as our CT sample seem to differ from teachers in other districts/subjects. While our teacher fixed-effects models effectively compare a teacher's performance in years in which they serve as CTs with performance in years in which they do not, in our preferred specification we use the full sample of teachers—including those teachers

Table 4

Effects of Serving as a Cooperating Teacher on Evaluation Metrics

	Observation Ratings		TVAAS—All Subjects		TVAAS—Math		TVAAS—ELA	
	(1) Fixed Effects	(2) Diff-in-Diff	(3) Fixed Effects	(4) Diff-in-Diff	(5) Fixed Effects	(6) Diff-in-Diff	(7) Fixed Effects	(8) Diff-in-Diff
Cooperating teacher	0.040***	0.053***	0.008	0.014*	−0.001	0.005	−0.003	0.005
	(0.007)	(0.007)	(0.006)	(0.006)	(0.012)	(0.013)	(0.007)	(0.007)
Ever cooperating teacher		0.108***		0.040***		0.075***		0.028***
		(0.007)		(0.006)		(0.011)		(0.005)
Mean outcome	3.885	3.885	0.063	0.040	0.133	0.081	0.037	0.021
Standard deviation	0.572	0.582	0.358	0.379	0.495	0.515	0.238	0.252
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher fixed effects	Yes	No	Yes	No	Yes	No	Yes	No
Experience fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
N	174,214	242,339	91,726	127,669	40,943	61,943	46,771	69,642
R^2	.771	.292	.689	.125	.702	.194	.619	.131
Adjusted R^2	.639	.286	.541	.110	.537	.167	.412	.105

Note. TVAAS = Teacher Value-Added Assessment System; ELA = English Language Arts. Robust standard error clustered by teacher in parentheses. Cooperating teacher is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a cooperating teacher. Experience is included as single indicators for Years 0 to 30 and as a pooled indicator for experience >30 years. We drop singleton observations from models with teacher fixed effects. This leads to different sample sizes between the fixed-effects and difference-in-differences (diff-in-diff) models.

$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

in non-CT blocks—to estimate coefficients for teaching experience and for the intercept term in a generalized difference-in-differences model. Thus, we wondered whether our estimates could be sensitive to our choice of analytic sample. To test this, we constrain our analyses to successively more restricted samples: (1) to teachers who teach in the same districts, same subject areas, and years as CTs in our sample; (2) to teachers who teach in the same school, subject areas, and years as CTs in our sample; and (3) only to teachers who served as CTs at least once. For all outcomes, the estimates for our preferred models have qualitatively similar estimates over the different estimation samples (see Appendix Table A1). However, the estimates for observation ratings decrease by about a quarter when we restrict the sample to teachers in the same blocks or same schools. This might be in line with our prior finding indicating that the blocks and schools where we observe CTs are different on baseline characteristics than other blocks and schools in the state. This will lead to a mechanical change in the coefficients for the covariates that we include in the model which could in part explain the difference in point estimates across the different sample specifications.[9]

Alternatively, this might indicate the presence of positive selection bias that is not fully accounted for in our preferred models but is accounted for in the models that restrict the sample to teachers in the same block or school. As an additional robustness check for our sample choice, we use a nearest neighbor matching algorithm to construct a sample of teachers who have observed characteristics similar to CTs but that were not picked to serve as CTs. We report these estimates in Table 5, alongside the estimates for the teacher fixed-effects and difference-in-differences models. Overall, we observe that the estimates for observation ratings have the same sign and magnitude across the different estimation models. Our matching algorithm matches on up to 2 years of prior evaluation scores; thus, we are matching CTs with non-CTs that have similar patterns of returns to experience preceding the CT years. While differences in pretrends could explain the positive effects on observation ratings in CT service years for our difference-in-differences specifications, these pretrends are unlikely to explain observed effects in our matched-sample models. Results for TVAAS appear to be significant and greater in magnitude for the matched-sample model. This might indicate that the matched-sample models, and to an extent the difference-in-differences models, could fail to account for self-selection bias. Said another way, teachers who were selected to be CTs could be different in unobserved ways from teachers who were not selected (e.g., stronger motivation). The teacher fixed-effects models account for these unobserved differences by leveraging the within-teacher variation in evaluation scores for CTs, assuming that these unobserved differences are constant within a teacher during our observation period. Failing to account for these unobserved differences could lead to estimates that are biased upward.

Table 5
Coefficient Sensitivity to Estimation Method

	(1) Observation Ratings	(2) TVAAS All Subjects	(3) TVAAS Math	(4) TVAAS ELA
Fixed-effects model	0.040***	0.008	−0.001	−0.003
Difference-in-differences model	0.053***	0.014*	0.005	0.005
Matched sample	0.059***	0.039**	0.092**	0.007

Note. ELA = English Language Arts; TVAAS = Teacher Value-Added Assessment System. This table reports the sensitivity of the cooperating teacher (CT) coefficient to various model specifications. The fixed-effects models include controls for years of experience, year and teacher fixed-effects. The difference-in-differences models include controls for years of experience, year and school fixed effects. The matched sample models report the Average Treatment Effects on the Treated (ATET) on teachers who serve as CTs. We fuzzy match using Mahalanobis distance on up to two prior years of evaluation data and years of experience at time of serving as a CT. We exact match on teacher background characteristics. We remove two CTs who do not match with other teachers in the state on background characteristics.
$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

Research Question 2: Are the Effects Different for Different Groups of CTs?

In this section, we investigate whether the effects of serving as a CT differ for different groups of teachers. We begin by examining heterogeneity for different quartiles of effectiveness. We then consider differences in estimates for teachers who work in different school levels.

Heterogeneity by Effectiveness Quartile

Table 6 reports the estimates that include an interaction term between the CT indicator and quartile indicator. We interpret the estimate for these interaction terms as the effect of serving as a CT for teachers in the various quartiles of effectiveness. We find positive effects of serving as a CT for teachers in all four quartiles of observation ratings. Moreover, we find that teachers in the lowest quartile benefit the most from serving as a CT compared with teachers in the other quartiles. A possible explanation for this pattern of results is that the ceiling effect built into the observation score rubric negatively biases the effects of serving as CT for teachers in the upper quartiles of effectiveness. In this case, the observation scores of more effective teachers do not have as much room for improvement as the scores of less effective teachers.[10] It is also possible that any contemporaneous professional development benefits of serving as a CT might affect lower performing CTs most.

Table 6

Heterogeneity by Quartile for Research Question 1

	(1) Observation Ratings	(2) TVAAS All Subjects	(3) TVAAS Math	(4) TVAAS ELA
Cooperating Teacher # Quartile 1	0.108*** (0.023)	0.001 (0.013)	−0.049+ (0.029)	−0.031 (0.024)
Cooperating Teacher # Quartile 2	0.026* (0.013)	−0.003 (0.009)	−0.028 (0.019)	−0.020+ (0.011)
Cooperating Teacher # Quartile 3	0.030** (0.011)	0.008 (0.008)	−0.004 (0.019)	0.004 (0.010)
Cooperating Teacher # Quartile 4	0.023** (0.009)	0.031* (0.014)	0.069* (0.027)	0.027+ (0.015)
Mean outcome	3.885	0.040	0.081	0.021
Standard deviation	0.582	0.379	0.515	0.252
Year fixed effects	Yes	Yes	Yes	Yes
Teacher fixed effects	Yes	Yes	Yes	Yes
Experience fixed effects	Yes	Yes	Yes	Yes
N	233,149	117,034	52,933	60,082
R^2	.748	.662	.671	.585
Adjusted R^2	.643	.527	.521	.398

Note. ELA = English Language Arts; TVAAS = Teacher Value-Added Assessment System. Robust standard error clustered by teacher in parentheses. Cooperating teacher is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a cooperating teacher. Experience is included as single indicators for Years 0 to 30 and as a pooled indicator for experience above 30 years. We drop singleton observations from models with teacher fixed effects. Quartiles are calculated for each outcome using the teacher fixed effect from a regression that includes an indicator for being a cooperating teacher, time-varying school characteristics, teacher fixed effects, and year fixed effects.

+$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Results for TVAAS tell a different story. We find that the effects of serving as a CT increase along the instructional effectiveness continuum. We observe positive effects on TVAAS scores only for teachers in the fourth quartile of effectiveness and possibly negative, but imprecisely estimated and nonsignificant, effects for teachers in the first and second quartiles of effectiveness. Finding effects to be more negative for lower performing teachers is consistent with what Goldhaber et al. (2018b) found in their sample of teachers from Washington state, where they hypothesize that more effective teachers are better able to buffer any potential negative effects of their mentees. However, while they found significant, negative effects overall (across quartiles), with the most negative effects concentrated in the lowest quartile, we find no significant effect overall (across quartiles) and instead small positive effects among teachers in the top quartile of instructional effectiveness.

One possibility is that our results are entirely driven by regression to the mean in evaluation scores (see Goldhaber et al., 2018b). In this case, we might conflate random year-to-year variation in evaluation scores with effects of serving as a CT. Specifically, if teachers' service (as CT) years coincide with years in which they also happen to randomly be at their peak performance, then they will tend to naturally regress to their average performance in post-CT years; this could lead to estimates like the ones that we observe for observation ratings in Table 6.[11] We check whether our results are sensitive to the way that we calculated the quartile of effectiveness by conducting a Monte Carlo simulation of the effects of serving as a CT on a placebo sample of CTs. Using the CT blocks described earlier, we randomly select 1,000 cohorts of teachers who were not actually selected as CTs during our observation period; these cohorts serve as a placebo for serving as a CT. We then calculate the effects of serving as a placebo CT for these 1,000 cohorts in order for us to test the extent to which our results are sensitive to regression to the mean. Appendix Table A3 shows the results from this Monte Carlo simulation. If regression to the mean were at play, then we would expect that the effects of regression to the mean on evaluation scores as a result of serving as a CT for the placebo group to have estimates similar to what we found for our CT sample. Instead, we find all the point estimates for the placebo CT sample are all close to zero and their 95% confidence intervals are centered at zero. More simply, we find no placebo effect of serving as a CT on evaluation scores for teachers along the effectiveness continuum. This suggests that our estimates for the effects of serving as a CT for each quartile of effectiveness are robust against teachers' evaluation scores regressing to the mean.

Heterogeneity by School Level

Table 7 displays the results for the effects of serving as a CT by school level. We find that the positive results on observation ratings are driven by

Table 7

Heterogeneity by School Level for Research Question 1

	(1) Observation Ratings	(2) TVAAS All Subjects	(3) TVAAS Math	(4) TVAAS ELA
Cooperating Teacher # Elementary School	0.052*** (0.009)	0.002 (0.008)	−0.017 (0.014)	−0.002 (0.011)
Cooperating Teacher # Middle School	0.032* (0.016)	0.002 (0.009)	−0.006 (0.022)	−0.004 (0.011)
Cooperating Teacher # High School	0.022 (0.014)	0.021 (0.016)	0.079 (0.058)	−0.016 (0.015)
Cooperating Teacher # Other School	0.031 (0.025)	0.033 (0.023)	0.009 (0.045)	0.016 (0.035)
Mean outcome	3.882	0.065	0.135	0.037
Standard deviation	0.570	0.356	0.494	0.238
Year fixed effects	Yes	Yes	Yes	Yes
Teacher fixed effects	Yes	Yes	Yes	Yes
Experience fixed effects	Yes	Yes	Yes	Yes
N	170,464	87,080	38,666	4,4243
R^2	.771	.689	.700	.618
Adjusted R^2	.639	.534	.527	.402

Note. ELA = English Language Arts; TVAAS = Teacher Value-Added Assessment System. Robust standard error clustered by teacher in parentheses. Cooperating teacher is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a cooperating teacher. Experience is included as single indicators for Years 0 to 30 and as a pooled indicator for experience >30 years. We drop singleton observations from models with teacher fixed effects. This leads to different sample sizes between the fixed effects and diff-in-diff models. Quartiles are calculated for each outcome using the teacher fixed effect from a regression that includes an indicator for being a cooperating teacher, time-varying school characteristics, teacher fixed effects, and year fixed effects.
$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

CTs who teach in elementary and middle schools and that the evaluation scores of teachers in high schools or other schools do not change when serving as a CT. Though we are not entirely sure why we observe these differences by school level, we explore possible explanations in the Discussion and Conclusion section. We also find that estimates for serving as a CT on TVAAS scores are mostly similar across the different school settings. However, the results seem to suggest that high school mathematics teachers' scores increase the year they serve as CTs but that these point estimates are imprecisely estimated.

Research Question 3: Do Teachers Perform Differently in Years After They Serve as CTs?

Based on our most conservative (i.e., teacher fixed effects) estimates, we find small and positive effects on observation ratings and null effects on TVAAS scores for years in which teachers serve as CTs. One possibility, though, is that the effects of serving as a CT are not immediate, but instead are observed in subsequent years. As discussed in our logic model, for example, if serving as a CT functions as a form of professional development, then we might not expect to observe increases in performance during the year a teacher serves, but perhaps in following years. In the next section, we turn to Research Question 3, where we estimate different effects for the years during which teachers serve as CTs and for years following that experience.

Table 8 reports the results of the teacher fixed-effects and difference-in-differences estimates of the effects of serving as a CT during postservice years. Specifically, we compare performance while serving as a CT and after serving as a CT with the evaluation scores during the time before serving as a CT. For observation ratings, CTs' evaluations do not increase, on average, in years following serving as a CT (see Columns 1 and 2). We note that the point estimates from the difference-in-differences model change sign but remain nonsignificant. That is, both specifications indicate that serving as a CT does not have a lasting impact on observation ratings beyond the years during which teachers serve as CTs.

On the other hand, results for TVAAS scores show that CTs' scores decline in the period after serving as a CT. TVAAS scores for the years in which teachers serve as CTs are similar to their scores for years before serving as a CT. However, scores in years after serving are lower than scores in years prior to serving. These results might highlight unobserved differences between CTs and non-CTs that we are not able to control in our main models. In fact, the negative effects for TVAAS scores disappear once we restrict our analyses to only teachers who ever serve as a CT (see Appendix Table A4, Column 3). This could suggest that CTs have differential returns to experience on TVAAS scores (this is supported by Atteberry et al., 2015, who find differential returns to experience by quartile of performance). Specifically,

Table 8

Effects of Serving as a Cooperating Teacher on Growth of Evaluation Metrics

	Observation Ratings		TVAAS—All Subjects		TVAAS—Math		TVAAS—ELA	
	(1) Teacher FE	(2) Diff-in-Diff	(3) Teacher FE	(4) Diff-in-Diff	(5) Teacher FE	(6) Diff-in-Diff	(7) Teacher FE	(8) Diff-in-Diff
Cooperating teacher	0.046***	0.049***	−0.002	−0.004	−0.010	−0.013	−0.011	0.001
	(0.009)	(0.009)	(0.007)	(0.008)	(0.015)	(0.015)	(0.008)	(0.008)
After cooperating teacher	0.015	−0.005	−0.020*	−0.030***	−0.016	−0.024	−0.017*	−0.007
	(0.009)	(0.010)	(0.008)	(0.009)	(0.017)	(0.016)	(0.008)	(0.008)
Ever cooperating teachers		0.112***		0.058***		0.091***		0.032***
		(0.009)		(0.008)		(0.014)		(0.006)
Mean outcome	3.885	3.885	0.040	0.040	0.081	0.063	0.021	0.021
Standard deviation	0.582	0.582	0.379	0.379	0.515	0.502	0.252	0.252
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher FE	Yes	No	Yes	No	Yes	No	Yes	No
Experience FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	No	Yes	No	Yes	No	Yes	No	Yes
N	233,149	242,339	117,034	127,669	52,933	61,943	60,082	69,642
R^2	.748	.292	.662	.125	.671	.194	.585	.131
Adjusted R^2	.643	.286	.527	.111	.521	.167	.398	.105

Note. ELA = English Language Arts; TVAAS = Teacher Value-Added Assessment System; FE = fixed effects. Robust standard error clustered by teacher in parentheses. Cooperating teacher is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a cooperating teacher. Experience is included as single indicators for Years 0 to 30 and as a pooled indicator for experience >30 years. We drop singleton observations from models with teacher fixed effects. This leads to different sample sizes between the fixed effects and (difference-in-differences) diff-in-diff models.

$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

CTs experience relatively higher growth on TVAAS in years leading up to service years. This performance may increase the likelihood that teachers are tapped to serve as CTs; given a bump in performance during years leading up to serving. A postserving decline may be expected if non-CTs close the TVAAS gap during the post-CT period.

Similar to Research Question 1, we explore whether our results are sensitive to sample selection. Appendix Table A4 presents the results for the teacher fixed-effects models on restricted samples of teachers. The estimate directions and magnitudes are similar across the estimation samples for the main effect of serving as a CT. The estimates for the period following serving as a CT appear to change somewhat depending on the sample that we use. For observation ratings, we find that the positive but insignificant estimate for the years following serving as a CT appears to move toward a null but imprecise estimate. For TVAAS scores, we find that the negative effect on the years following serving as a CT appears to move toward zero.[12]

Discussion and Conclusion

There is growing evidence that recruiting more instructionally effective teachers to serve as CTs is a promising approach to improving the preparation that teacher candidates receive and, subsequently, the instructional effectiveness of the incoming supply of new teachers. So why are program leaders reporting that it can be difficult to get our most instructionally effective teachers to serve as CTs? The challenge appears to be multifaceted, and this study investigates one factor: that teachers are hesitant to mentor a teacher candidate for fear that they may receive lower evaluations. Our results suggest that any concerns over declining evaluations are not warranted. Rather, we find observation ratings may even increase while student achievement gains are unaffected. The implications are that instructionally effective teachers who are considering becoming CTs should not let fears over evaluation scores deter them from serving. Moreover, program and district leaders charged with recruiting these teachers to serve can assure potential CTs that such fears are likely unjustified.

Since this article is the first to examine effects on observation ratings, a unique contribution is finding that teachers' concurrent observation ratings may actually benefit by mentoring a candidate. Regarding student achievement gains, our main results are similar to those of Goldhaber et al. (2018b). Both studies found no effects of serving as a CT on average student achievement gains, though our study found small, positive effects in alternate model specifications.

Goldhaber et al. (2018b) also found that the effects of serving as a CT on math achievement were negative among lowest performing teachers, but effectively zero for other quartiles. They point out that this finding is somewhat surprising in light of prior evidence that principals sometimes place

candidates in classrooms of less effective teachers to help boost performance in those classrooms; in fact, the authors find the reverse to be true—placing candidates in these classrooms appears to harm performance. They conclude that more instructionally effective teachers are likely better able to buffer against the negative effects of mentoring a teacher candidate. In comparison, we find that the coefficients for serving as a CT decrease as CT effectiveness decreases. However, we find positive and significant effects for teachers in the top quartile, and negatively trending but nonsignificant effects for teachers in lower quartiles. These results do not seem to be consistent with an explanation that higher performing teachers are mitigating the negative effects of mentoring a candidate; rather, our results seem to suggest that higher performing teachers may actually benefit from mentoring a candidate.

In terms of observation ratings, we find that all teachers, across all quartiles of prior performance, receive significantly higher ratings in years that they serve as CTs. However, we find that lowest quartile CTs tend to benefit most.[13] One might be tempted to conclude, based on this latter finding, that program and school leaders should place candidates with lower performing teachers. However, such a conclusion, we believe, would likely be premature in light of other evidence. First, there is strong evidence that graduates have better early career performance when they learn to teach with more instructionally effective CTs (Goldhaber et al., 2018b; Ronfeldt, Brockman & Campbell, 2018; Ronfeldt, Matsko, et al., 2018). Second, results from our study and Goldhaber et al. (2018b) indicate that student achievement gains among lower performing teachers tend to decline in years they mentor candidates. Finally, as we describe above, we do not actually know if the overall teaching and learning quality are benefitting in classrooms of lowest performing teachers who mentor candidates or if teachers' new roles as CTs somehow change their evaluation procedures in ways that benefit their ratings without necessarily benefitting their teaching (see Figure 1 Perceived Changes). We encourage future work to examine the specific mechanisms by which teachers' evaluations change as a result of mentoring candidates, including why there appear to be differences by the level of instructional effectiveness of CTs.

Prior research has established that recruiting the most instructionally effective teachers to serve as CTs is likely to benefit the new supply of prospective teachers and those schools and districts that hire them. Our present study suggests that this strategy is likely also to benefit those teachers who serve as CTs, or at least cause the least harm. Though there are some subtle differences between our results and those of Goldhaber et al. (2018b) and SAS Institute (2014), a common conclusion from all three studies is that the student achievement gains of the least instructionally effective teachers are likely to suffer when they mentor candidates, whereas the achievement gains and observation ratings of the most instructionally effective teachers

are likely to go unchanged and possibly even improve. Taken together, the existing research then tends to suggest that placing candidates with the most instructionally effective teachers is likely to result in the greatest overall benefit and least harm.

A limitation of our study is that we do not have comprehensive data identifying all teachers who served as CTs across the state and across the years included in our study. Rather, our CT data come from only those TEPs that kept and were willing to share these data and only for years that were included in their records. Thus, our coverage across TEPs and years is uneven, and our results are not necessarily generalizable across the state. It is possible that the CTs for the particular TEPs and years in our sample may respond differently to serving than the CTs we do not observe. Though unlikely, it is possible that teachers in our sample improved on observation ratings when they served but teachers outside our sample declined in performance. This might occur, for example, if CTs in unobserved programs may have had different motivations for serving than those in the programs we observed. The study by Goldhaber et al. (2018b) also did not have full coverage of programs in Washington state and so may be subject to similar limitations. We are currently in negotiations with the TDOE to see if we can access comprehensive data on CTs across all programs in Tennessee for future cohorts.

More research is needed to understand possible mechanisms by which teachers get a boost in observation ratings during the years in which they serve as a CT. As identified in our logic model, one possibility is that serving as a CT does indeed boost the quality of instruction. It might be, for example, that, in years they are serving as CTs, having an additional adult in the classroom helps with instruction by increasing the amount of individual instructional time each student has with a teacher. In years they serve as CTs, teachers also might invest more in instructional planning as a result of needing to onboard another teacher and ensure they are modeling good practice.

Another possibility is that teachers who serve as CTs must schedule evaluations on days or during sections/periods when their candidates are not lead teaching. This likely means that unscheduled observations (for evaluation) are less common in years that teachers mentor candidates. It also could mean that CTs are able to be more strategic about when they schedule observations/evaluations—for example, during easier periods/classes or during subjects in which they especially excel—thus, reducing the impact of unannounced observations. In these ways, teachers could effectively boost their evaluations, possibly explaining the bumps in performance we observe during years they serve as CTs. These explanations are consistent with finding that lowest quartile teachers benefit most on observation ratings when serving as CTs, as one might expect strategically planned evaluations to benefit less-effective teachers most. They are also consistent with finding little to no

effects of serving as a CT on TVAAS scores, where student achievement, rather than scheduled observations by raters, dictate performance.

One additional consideration is that we find the positive effects of serving as a CT on observation ratings to be concentrated in elementary schools. There are many reasons why this might be true. It could be, for example, that elementary candidates are better able to form personal relationships with students because they are with them all day or because younger students are more willing to connect with new teachers in their classrooms; this, in turn, likely translates into a stronger instructional environment. If this were the case, though, we would expect teachers' TVAAS, and not just observation ratings, to increase in years they serve as CTs.

Alternatively, one of the main differences between elementary and secondary teachers is that the former are tasked with teaching all subjects, even ones in which they are less knowledgeable or effective. It is possible that elementary teachers who mentor candidates are more inclined to hand over lead teaching responsibilities in subjects in which they feel less proficient. If so, this could result in evaluators being more likely to evaluate CTs when teaching their stronger subjects and, thus, to rate them higher than in other years.[14] If this were the case, then we would expect mentoring a candidate to likely benefit lower performing elementary CTs the most, which is what we observe (see Appendix Table A6).

It is true that secondary teachers also often teach multiple courses. Secondary teachers could also then assign their teacher candidates to the subjects that are their weakest. However, we believe that secondary candidates are often more specialized in their subject matter/content focus and more likely to request a specific class that is a match. Compared with elementary teachers, this would likely place more constraints on secondary CTs in terms of which parts of the school day that they would be able to hand over lead teaching responsibilities to candidates, that is, secondary CTs may have somewhat less flexibility than elementary CTs in how they assign their candidates.

An important next step for future research is to interrogate the mechanisms by which observation ratings increase in years that teachers serve as CTs and, relatedly, whether the increases reflect improvements in actual teaching quality or instead changes in evaluation processes that result in better evaluations but not necessarily better teaching. If boosted performance among CTs is explained by having opportunities to game the evaluation system, whether intentional or unintentional, then we expect that some will argue that it seems inequitable for teachers who mentor a candidate to gain such an advantage. Though we understand this perspective, we also recognize that mentoring a candidate is a tremendous amount of additional work for classroom teachers and is often unrecognized, underappreciated, and not rewarded. One recent study found that CTs typically received only about $300 for mentoring a candidate, and that many were not

compensated at all (Matsko et al., 2019). Especially given that CTs can have meaningful, positive impacts on the instructional effectiveness of the incoming supply of teachers, they may be deserving of advantages during the years in which they mentor. In fact, one consideration might be to relieve teachers of being evaluated in the years in which they mentor a candidate. Doing so would remove any concerns about CTs gaming the evaluation system and would offset fears that mentoring a candidate might harm evaluation scores—even though our results suggest that these fears may be unwarranted—while providing a low-cost incentive to serve.

Finally, our research extends prior work by testing whether or not serving as a CT has a longer term effect on evaluation during post-CT years. A positive effect in post-CT years could suggest that mentoring serves a professional development function. For observation ratings, we find post-CT performance to decline back to pre-CT levels. For TVAAS, we find post-CT performance to actually be somewhat worse than pre-CT levels. However, when we constrain models only to individuals who ever served as CTs, the postserving estimates are similar to pre-CT estimates; this may suggest that CTs are not actually doing worse in post-CT years, but that instead, non-CTs tend to have stronger relative returns. Either way, while we find some boost to evaluations during the years in which teachers serve as CTs, we find no evidence that serving as a CT makes individuals better teachers in post-CT years. Thus, serving as a CT does not appear to function as a form of long-term professional development. These results differ from Goldhaber et al. (2018b), who find student achievement gains to increase significantly in post-CT years and conclude that serving as a CT may serve as a form of professional development for teachers.

More research needs to investigate why these results in Washington differ from ours in Tennessee. One possibility is that the different labor markets, policy, and evaluation contexts afford and constrain different responses by teachers who serve as CTs. Another possibility is that the differences result from different ways that the two studies measured value-added to student achievement gains (VAMs) and modeled effects on achievement gains. In particular, in their construction of VAMs, Goldhaber and colleagues used student-level data and controlled for many student characteristics, whereas our study, due to our data sharing agreements, depends on teacher-level TVAAS measures which do not adjust for the same covariates. If such differences in VAM construction were responsible for differences in effects on post-CT outcomes, then we would have likely also expected differences in concurrent effects, but our results are similar.

While more research is clearly needed to understand the mechanisms by which serving as a CT impact evaluations—during concurrent and post-CT years—the main policy conclusion from this study is generally promising and consistent with conclusions from prior research: Serving as a CT does not appear to harm a teacher's concurrent or future performance evaluations

and may even be of benefit. For teacher education program leaders, a related implication is that they should not be concerned about potential unintended consequences to teachers they recruit to serve as CTs. In fact, sharing that there may even be some evaluation benefits could assist in their recruitment efforts.

Appendix

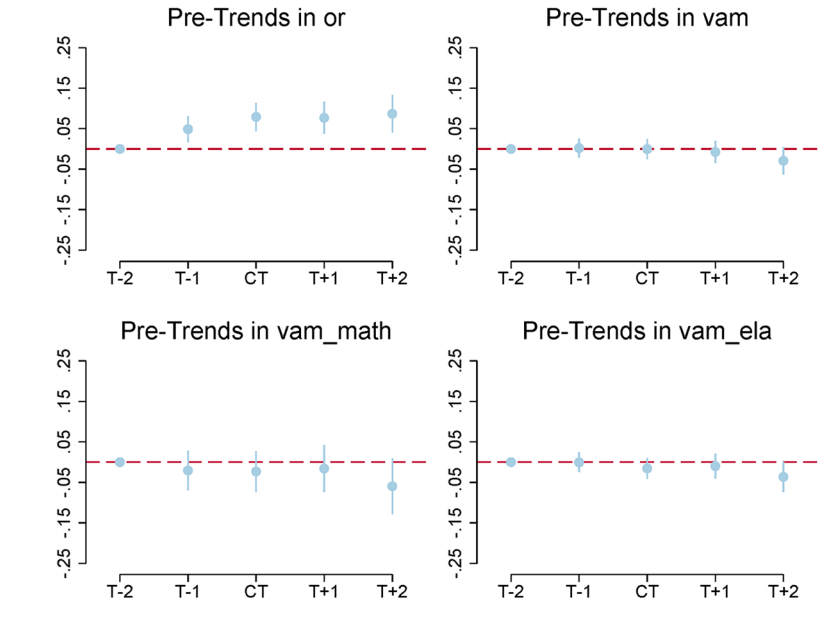

Figure A1. **Event study of serving as a cooperating teacher (CT) on outcomes of interest.**

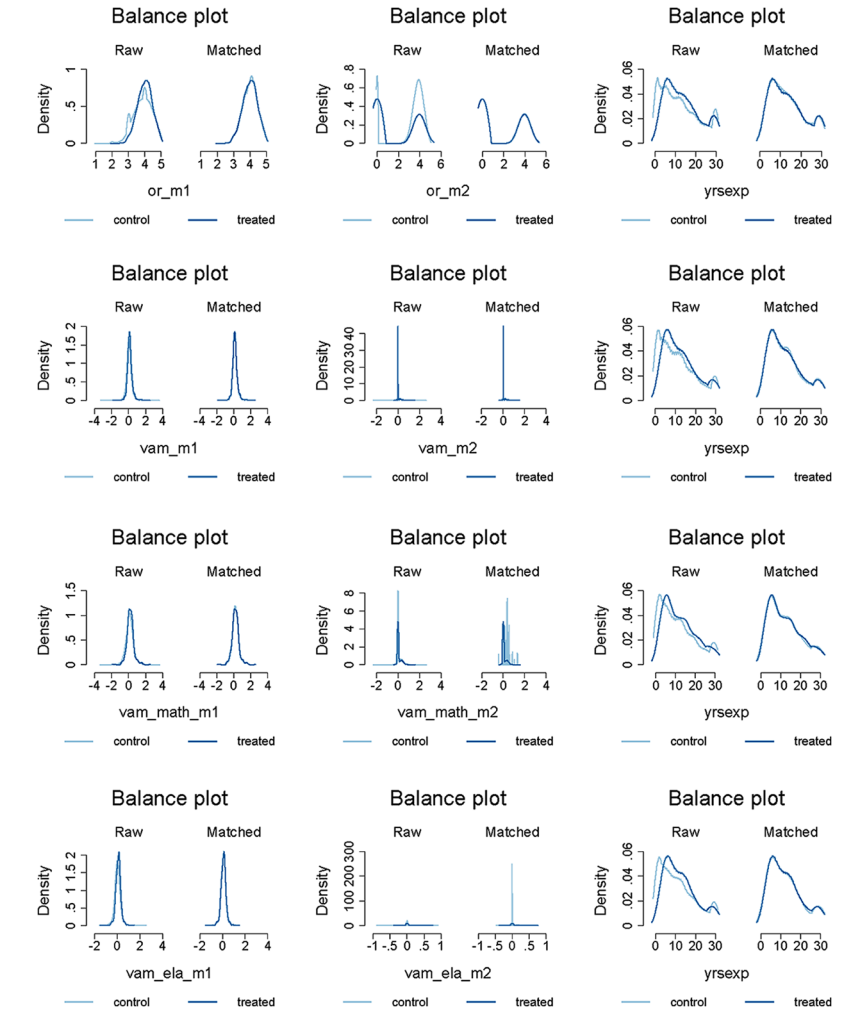

Figure A2. **Density distribution of fuzzy matched covariates.**

Table A1

Robustness Checks of Teacher Fixed-Effects Models on Different Subsamples of Teachers

	(1) Main Effects	(2) CT Blocks	(3) CT Schools	(4) CT Ever
Panel A: Observation ratings				
CT	0.040*** (0.007)	0.028*** (0.007)	0.030** (0.009)	0.035*** (0.007)
Mean outcome	3.870	3.868	3.878	4.022
Standard deviation	0.578	0.570	0.534	0.485
Year fixed effects	Yes	Yes	Yes	Yes
Teacher fixed effects	Yes	Yes	Yes	Yes
Experience fixed effects	Yes	Yes	Yes	Yes
N	174214	81513	19880	10127
R^2	.771	.784	.816	.741
Adjusted R^2	.639	.638	.671	.603
Panel B: TVAAS—All subjects				
CT	0.008 (0.006)	0.008 (0.006)	0.008 (0.008)	0.007 (0.006)
Mean outcome	0.063	0.068	0.077	0.113
Standard deviation	0.358	0.326	0.309	0.305
Year fixed effects	Yes	Yes	Yes	Yes
Teacher fixed effects	Yes	Yes	Yes	Yes
Experience fixed effects	Yes	Yes	Yes	Yes
N	91,726	52,838	11,817	7,042
R^2	.689	.709	.728	.668
Adjusted R^2	.541	.537	.522	.517

Note. CT = cooperating teacher; TVAAS = Teacher Value-Added Assessment System. Robust standard error clustered by teacher in parentheses. CT is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a CT. Experience is included as single indicators for Years 0 to 30 and as a pooled indicator for experience above 30 years. Models (1) and (4) include singleton observations by teacher-year. This leads to different sample sizes between the fixed-effects and difference-in-differences (diff-in-diff) models.
$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

Table A2
Nearest Neighbor Matching Quality

	Mean		Variance	
	Raw	Matched Sample	Raw	Matched Sample
Panel A: Observation ratings (OR)				
OR—2 years prior	−0.847	0.000	1.874	1.000
OR—1 year prior	0.271	0.002	0.613	1.004
Years of experience	0.175	0.016	0.849	1.017
N	45,220	4,608		
Panel B: TVAAS				
TVAAS—2 years prior	−0.097	0.012	0.503	1.135
TVAAS—1 year prior	0.255	0.013	0.777	1.180
Years of experience	0.181	0.008	0.866	0.987
N	18,424	2,236		
Panel C: TVAAS Mathematics				
TVAAS Math—2 years prior	−0.127	0.015	0.427	1.088
TVAAS Math—1 year prior	0.186	0.007	0.757	1.122
Years of experience	0.172	0.021	0.865	0.987
N	8,753	988		
Panel D: TVAAS ELA				
TVAAS ELA—2 years prior	−0.036	0.009	0.704	1.081
TVAAS ELA—1 year prior	0.265	0.008	0.823	1.126
Years of experience	0.162	0.016	0.817	0.969
N	9,684	1,176		

Note. ELA = English Language Arts; TVAAS = Teacher Value-Added Assessment System. This table reports the standardized difference between cooperating teachers (CTs) and non-CTs on the variables we used to construct the nearest neighbor matched sample. Values close to 0 for the matched sample means and close to 1 for the matched sample variance indicate that the matching procedure was able to identify a non-CT sample similar to the observed CT sample.

Table A3

Monte Carlo Simulation

	(1) Observation Ratings		(2) TVAAS—All Subjects		(3) TVAAS—Math		(4) TVAAS—ELA	
Placebo CT # Q1	–0.003	[–0.031, 0.027]	–0.006	[–0.033, 0.024]	0.004	[–0.058, 0.063]	–0.019	[–0.046, 0.009]
Placebo CT # Q2	–0.008	[–0.032, 0.016]	–0.006	[–0.019, 0.009]	0.001	[–0.030, 0.030]	–0.012	[–0.029, 0.004]
Placebo CT # Q3	–0.004	[–0.024, 0.017]	0.002	[–0.011, 0.015]	0.010	[–0.017, 0.039]	–0.004	[–0.021, 0.012]
Placebo CT # Q4	0.002	[–0.016, 0.020]	0.006	[–0.015, 0.028]	0.007	[–0.033, 0.047]	–0.005	[–0.030, 0.020]

Note. ELA = English Language Arts; TVAAS = Teacher Value-Added Assessment System. This table reports the results of a Monte Carlo simulation that draws 1,000 placebo CTs and calculates the placebo effect of serving as a cooperating teacher (CT) on evaluation scores. The model adjusts for year fixed effects. Quartiles are calculated using the method that we used to estimate the heterogeneity by quartile using evaluation years 2011–2015. 95% credible intervals are in brackets.
$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

Table A4

Robustness Checks of Growth Models on Different Subsamples of Teachers

	(1) Main Effects	(2) CT Blocks	(3) CT Schools	(4) CT Ever
Panel A: Observation ratings				
CT	0.046*** (0.009)	0.040*** (0.010)	0.034* (0.013)	0.032** (0.011)
Years following serving as a CT	0.015 (0.009)	0.027+ (0.014)	0.007 (0.019)	−0.003 (0.016)
Mean outcome	3.885	3.847	3.841	4.038
Standard deviation	0.582	0.582	0.561	0.497
Year fixed effects	Yes	Yes	Yes	Yes
Teacher fixed effects	Yes	Yes	Yes	Yes
Experience fixed effects	Yes	Yes	Yes	Yes
N	233,149	81,807	19,880	13,157
R^2	.748	.784	.816	.707
Adjusted R^2	.643	.638	.671	.599
Panel B: TVAAS—All subjects				
CT	−0.002 (0.007)	−0.001 (0.008)	0.009 (0.011)	0.010 (0.009)
Years following serving as a CT	−0.020* (0.008)	−0.018+ (0.011)	0.005 (0.016)	0.008 (0.013)
Mean outcome	0.040	0.048	0.046	0.093
Standard deviation	0.379	0.349	0.330	0.314
Year fixed effects	Yes	Yes	Yes	Yes
Teacher fixed effects	Yes	Yes	Yes	Yes
Experience fixed effects	Yes	Yes	Yes	Yes
N	117,034	52,918	11,817	8,468
R^2	.662	.709	.728	.645
Adjusted R^2	.527	.537	.522	.515

Note. CT = cooperating teacher; TVAAS = Teacher Value-Added Assessment System. Robust standard error clustered by teacher in parentheses. CT is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a CT. Experience is included as single indicators for Years 0 to 30 and as a pooled indicator for experience above 30 years. In addition, experience variable is interacted with the years following indicator, allowing differential returns to experience after a teacher first serves as a CT. Models (1) and (4) include singleton observations by teacher-year.

+$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table A5

Two-Stage Estimation for Cooperating Teacher Growth Trajectories

	Observation Ratings		TVAAS All Subjects		TVAAS Math		TVAAS ELA	
	(1) Teacher FE	(2) Diff-in-Diff	(3) Teacher FE	(4) Diff-in-Diff	(5) Teacher FE	(6) Diff-in-Diff	(7) Teacher FE	(8) Diff-in-Diff
Cooperating teacher	.051	.048	−.002	−.011	−.015	−.015	−.012	−.008
After cooperating teacher	.023	−.005	−.021	−.043	−.025	−.038	−.020	−.027
Ever cooperating teachers		.112		.066		.098		.043

Note. ELA = English Language Arts; TVAAS = Teacher Value-Added Assessment System; FE = fixed effects. Standard errors are not reported because they are not calculated for the two-stage Papay-Kraft estimation correction. Cooperating teacher is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a cooperating teacher. Experience is included as single indicators for Years 0 to 30 and as a pooled indicator for experience above 30 years. We drop singleton observations from models with teacher fixed effects. This leads to different sample sizes between the fixed-effects and difference-in-differences (diff-in-diff) models.

Table A6

Heterogeneity by School Type and Quartile

	Observation Ratings				TVAAS—All Subjects			
	(1) Elementary School	(2) Middle School	(3) High School	(4) Other School	(5) Elementary School	(6) Middle School	(7) High School	(8) Other School
CT # Quartile 1	0.153*** (0.030)	0.012 (0.018)	0.051 (0.035)	0.032 (0.025)	0.003 (0.020)	−0.011 (0.012)	−0.007 (0.013)	−0.008 (0.027)
CT # Quartile 2	0.058 (0.051)	0.025+ (0.014)	0.041 (0.026)	0.027 (0.025)	0.081* (0.037)	0.004 (0.012)	0.002 (0.013)	0.028 (0.021)
CT # Quartile 3	0.089* (0.037)	0.035** (0.012)	0.029 (0.024)	−0.005 (0.019)	0.009 (0.038)	0.010 (0.018)	0.005 (0.024)	0.088** (0.034)
CT # Quartile 4	0.006 (0.027)	0.104* (0.048)	0.042 (0.050)	−0.020 (0.109)	0.036 (0.051)	0.018 (0.018)	−0.039 (0.031)	0.077 (0.052)
N	228,417				112,476			
R^2	.749				.663			
Adjusted R^2	.644				.523			

Note. TVAAS = Teacher Value-Added Assessment System. Robust standard error clustered by teacher in parentheses. Cooperating teacher is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a cooperating teacher. Experience is included as single indicators for Years 0 to 30 and as a pooled indicator for experience above 30 years. We drop singleton observations from models with teacher fixed effects. This leads to different sample sizes between the fixed-effects and difference-in-differences (diff-in-diff) models. Quartile are calculated for each outcome using the teacher fixed effect from a regression that includes an indicator for being a cooperating teacher, time-varying school characteristics, teacher fixed effects, and year fixed effects.
$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

Notes

We appreciate the generous financial support that was provided for this research by the Institute of Education Sciences, U.S. Department of Education through the Statewide, Longitudinal Data Systems Grant (PR/Award R372A150015). Stacey Brockman, Emanuele Bardelli, and Hannah Mullman also received predoctoral support from the Institute of Education Sciences, U.S. Department of Education (PR/Award R305B150012). We also appreciate comments from John Papay, Dan Goldhaber, James Cowan, Kevin Schaaf, reviewers at the Tennessee Education Research Alliance on earlier drafts of this article, and from seminar attendees at our Causal Inference in Education Research Seminar presentation at the University of Michigan and conference attendees at the 2019 American Education Finance and Policy conference in Kansas City, MO. This project would not have been possible without the partnership, support, and data provided by the Tennessee Department of Education. Please note that the views expressed are those of the authors and do not necessarily reflect those of this study's sponsors, the Tennessee Department of Education, or the institutions to which the authors are affiliated.

[1]In this article, we use measures from Tennessee's educator evaluation system as our outcomes of interest. While this evaluation system is designed to capture multiple aspects of teaching, we are also aware that these measures could leave out important dimensions of teaching practice that are not addressed on standardized tests or observation rubrics. We want to stress that we are measuring the impacts of mentoring a candidate on these evaluation scores rather than changes in CTs' skills while mentoring candidates.

[2]The Tennessee Department of Education asked all educator preparation programs in the state to share their placement data for this project, including cooperating teacher-teacher candidate match data that is not currently collected as part of the teacher licensing process. Seventeen programs agreed to share their data. These data cover about 40% of the teacher candidates prepared in Tennessee during our period of observation.

[3]In most of our models, we restrict the evaluation data to cover the same time span as the CT data set. As a robustness check, we use the full evaluation data. Our results are robust against the data set that we use to estimate the effects of serving as a CT on evaluation scores.

[4]About 20% of Tennessee teachers are assessed using other rubrics than the Tennessee educator acceleration model rubric. As part of the state-wide educator assessment system, the Tennessee department of education calculates equated scores among the Tennessee educator acceleration model rubric scores and these other observational rubrics. We use these equated scores in our analyses.

[5]It is common for TEPs to ask candidates for their preferences in terms of districts in which they are willing to complete their student teaching and to then select placements in the requested districts (Krieg et al., 2016; Maier & Youngs, 2009). One reason for this is that candidates often have geographic and travel constraints.

[6]Goldhaber et al. (2018b) also found evidence of regression to the mean in their sample. We test for this issue using a Monte Carlo simulation described below. We do not find evidence that evaluation scores regress to the mean in our sample. This fact could be due to differences in the way that we calculated the effectiveness quartile for teachers and the way in which Tennessee calculates teacher value-added scores. First, we calculate quartile of effectiveness using all evaluation data available for each teacher. This is because we do not have access to evaluation data for the period preceding serving as a cooperating teacher. Second, TVAAS models differ from traditional value-added models insofar that scores for each teacher are calculated separately for each student cohort and that teacher value-added are calculated using empirical Bayes's estimates (Vosters et al., 2018).

[7]We observe that 83% of CTs are reported to serve only once during our observation period, 14% of CTs serve twice, 3% serve three times or more.

[8]As we discussed in the "Method" section, these results rely on a different set of assumptions than the teacher fixed-effects estimates. Namely, we are assuming that the evaluation scores for CTs and non-CTs follow parallel trends during the pre-CT period. The difference in results between the two specifications could suggest that this assumption is not met, that is, CTs have different returns to experience than non-CTs. While our analysis of parallel trends is partial, we find some potential evidence that pretrends are not parallel for observation ratings (see the "Method" section and Appendix Figure

A2). However, the results that we report in Appendix Tables A1 and A3 seem to suggest that that the estimates from our preferred models are well specified.

[9]To test whether the change in covariate coefficients could explain our results, we adjust the year fixed effects using the method that Papay and Kraft (2015) describe. These models' results are qualitatively identical to our preferred model estimates.

[10]Observation scores averages are 3.36 (*SD* = 0.36) for CTs in Quartile 1, 3.73 (*SD* = 0.30) for teachers in Quartile 2, 4.01 (*SD* = 0.28) for Quartile 3, and 4.39 (*SD* = 0.38) for Quartile 4.

[11]In detail, regression to the mean happens when a variable is measured with error. Point estimates from a regression model will include both the true effect and the effect of random measurement error. Since the effect of the random measurement error changes year-to-year, the point estimate for the effect of interest will also vary around its true point estimate. In our case, regression to the mean could explain the improvement in observation scores that we observe for teachers in the lower quartiles of effectiveness by suggesting that our CT estimates are based on a "good evaluation" (i.e., positive measurement error) year and that these teachers' observation ratings regress back to their mean for years following serving as CT.

[12]A possible explanation for these unstable estimates could be collinearity between the CT and following CT indicators, the experience fixed effects, and the years fixed effects. This would lead to unstable point estimates that are sensitive to the estimation sample that we use to identify the main effects. To address this concern, we use the two-stage adjustment strategy for year fixed-effects described in Papay and Kraft (2015). We first estimate the year fixed effect using a model that does not include teacher fixed effects. We then use the year-specific coefficients estimated in Stage 1 in our preferred model. The results from these models are consistent with the estimates from our preferred models (see Appendix Table A5), confirming a null effect on observation ratings for years following serving as a CT and a possible small and negative effect on TVAAS scores.

[13]It is important to point out that the lowest performing teachers in our CT sample are actually still more instructionally effective than the average teacher in the state.

[14]It also may be more difficult for evaluators to do unplanned observations and evaluations of elementary CTs, since they are more likely than secondary CTs to be with a student teacher throughout the entire day. Since teachers can prepare for planned observations, this may effectively boost evaluations for elementary CTs.

References

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, *1*(4). doi:10.1177/2332858415607834

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*, 416–440.

Campbell, S. L. (2014). *Quality teachers wanted: An examination of standards-based evaluation systems and school staffing practices in North Carolina middle schools* (Order No. 3633946). Available from ProQuest Dissertations & Theses A&I (1612601875).

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*. Advanced online publication. doi:10.3102/0002831218776216

Cantrell, S., & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study* (MET Project Research Paper). Retrieved from https://files.eric.ed.gov/fulltext/ED540958.pdf

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, 181–199.

Gitomer, D., & Bell, C. (2013). Evaluating teachers and teaching. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology. Vol. 3: Testing and assessment in school psychology and education* (pp. 415–444). Washington, DC: American Psychological Association.

Goldhaber, D., Krieg, J., & Theobald, R. (2018a). *Effective like me? Does having a more productive mentor improve the productivity of mentees?* (CALDER Working Paper No. 208-1118-1). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Goldhaber, D., Krieg, J., & Theobald, R. (2018b). *The costs of mentorship? Exploring student teaching placements and their impact on student achievement* (CALDER Working Paper). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

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

Grossman, P., Hammerness, K. M., McDonald, M., & Ronfeldt, M. (2008). Constructing coherence: Structural predictors of perceptions of coherence in NYC teacher education programs. *Journal of Teacher Education*, *59*, 273–287.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added. *American Educational Research Journal*, *51*, 73–112.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*, 794–831.

Jiang, J. Y., & Sporte, S. (2016). *Teacher evaluation in Chicago: Differences in observation and value-added scores by teacher, student, and school characteristics*. Retrieved from https://consortium.uchicago.edu/sites/default/files/2018-10/Teacher%20Evaluation%20in%20Chicago-Jan2016-Consortium.pdf

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

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

Loeb, S., & Candelaria, C. (2012). How stable are value-added estimates across years, subjects, and student groups? *Carnegie Knowledge Network*. Retrieved from https://cepa.stanford.edu/content/how-stable-are-value-added-estimates-across-years-subjects-and-student-groups

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

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty. *Educational Researcher*, *43*, 304–316.

Matsko, K. K., Ronfeldt, M., & Greene Nolan, H. (2019). *How different are they? Comparing preparation offered by traditional, alternative, and residency pathways*. Manuscript submitted for publication.

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

Mullman, H., & Ronfeldt, M. (2019). *The landscape of clinical preparation in Tennessee*. Manuscript submitted for publication.

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.

Ronfeldt, M. (2012). Where should candidates learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis*, *34*, 3–26.

Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education*, *66*, 304–320.

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher*, *47*, 405–418. doi:10.31102/0013189X18782906

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

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

Ronfeldt, M., Schwartz, N., & Jacob, B. (2014). Does pre-service preparation matter? Examining an old question in new ways. *Teachers College Record*, *116*(10), 1–46.

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

Shanks, R. (2017). Mentoring beginning teachers: Professional learning for mentees and mentors. *International Journal of Mentoring and Coaching in Education*, *6*(3), 158–163.

Spencer, T. L. (2007). Cooperating teaching as a professional development activity. *Journal of Personnel Evaluation in Education*, *20*, 211–226.

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*, 293–317.

St. John, E., Goldhaber, D., Krieg, J., & Theobald, R. (2018). *How the match gets made: Exploring candidate placements across teacher education programs, districts, and schools* (CALDER Working Paper). Retrieved from https://caldercenter.org/sites/default/files/CALDER%20WP%20204-1018-1.pdf

Vosters, K. N., Guranio, C. M., & Wooldridge, J. M. (2018). Understanding and evaluating the SAS® EVAAS® univariate response model (URM) for measuring teacher effectiveness. *Economics of Education Review*, *66*, 191-205. doi:10.1016/j.econedurev.2018.08.006

White, M. (2018). *Generalizability of observation instruments* (Unpublished doctoral dissertation). University of Michigan, Ann Arbor.

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014, May 13). *Evaluating teachers with classroom observations*. Washington, DC: Brown Center on Education Policy, Brookings Institute.

Manuscript received November 16, 2018
Final revision received August 7, 2019
Accepted August 8, 2019