



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

Matthew Ronfeldt
University of Michigan

Emanuele Bardelli
University of Michigan

Stacey Brockman
University of Michigan

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 negatively impact district evaluation scores. Using a unique dataset on over 4,500 CTs, we compare evaluation scores during years these teachers served as CTs to 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.

VERSION: May 2019

Suggested citation: Ronfeldt, M., Bardelli, E., Brockman, S., & Mullman, H. (2019). Will Mentoring a Student Teacher Harm My Evaluation Scores? Effects of Serving as a Cooperating Teacher on Evaluation Metrics (EdWorkingPaper No.19-69). Retrieved from Annenberg Institute at Brown University: <http://edworkingpapers.com/ai19-69>

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

Matthew Ronfeldt

Emanuele Bardelli

Stacey Brockman

Hannah Mullman

University of Michigan

Abstract: 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 negatively impact district evaluation scores. Using a unique dataset on over 4,500 CTs, we compare evaluation scores during years these teachers served as CTs to 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.

Acknowledgments: We appreciate the generous financial support that was provided for this research by the Institute of Education Sciences (IES), U.S. Department of Education through the Statewide, Longitudinal Data Systems Grant (PR/Award R372A150015). Stacey Brockman, Emanuele Bardelli, and Hannah Mullman also received pre-doctoral support from the Institute of Education Sciences (IES), U.S. Department of Education (PR/Award R305B150012). We also appreciate comments from John Papay, Dan Goldhaber, James Cowan, reviewers at the Tennessee Education Research Alliance on earlier drafts of this paper, and from seminar attendees at our Causal Inference in Education Research Seminar (CIERS) 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.

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; Ronfeldt, 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, & Reiningger, 2018).

Yet, many teacher education program leaders and state policymakers suggest that, despite their best efforts, teacher candidates are often placed with CTs who are 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 a candidate on student achievement, though lower-performing teachers have worse achievement in math.

More studies in different labor markets and policy environments are needed – like the present study in Tennessee – in order to test if 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 impacts 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, 2008). 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 to 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 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 non-significant) effects for bottom-quartile teachers; this is somewhat inconsistent with Goldhaber and colleagues (2018), 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. 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 by 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 even increase evaluations. An implication is that teacher education programs and policymakers should strive to recruit instructionally effective teachers to serve as CTs and that these teachers should consider serving as

CTs because, beyond benefiting the next generation of teachers, doing so and may even improve their own evaluations.

Literature Review

The vast majority of existing literature on CTs focuses on the effects of CTs on teacher candidates, whereas this study investigates the effects of supervising a candidate on CTs; however, there are currently no published, peer-reviewed 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, 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, preliminary 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 (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 to peers whose CTs received ratings of 3.0. They also found the student achievement value-added scores of candidates and their CTs to be significantly and positively correlated. Likewise, in a study using data from Chicago Public Schools (CPS), Ronfeldt, Matsko, Greene Nolan et al. (2018), found that each additional point on a CTs' observational ratings (on a scale of 1-4) was associated with a 0.16 point gain for their preservice candidates' observational

rating during their first year, an amount that is comparable to the average difference on observation ratings between teachers in their first year and teachers with between two and five years of experience in Chicago (Jiang & Sporte, 2016). More recently in Washington, Goldhaber, Krieg, and Theobald (2018) have also found strong, positive associations between the math achievement of mentees and mentors, and more modest, but still positive, associations in ELA.

CT recruitment and selection

The literature reviewed thus far suggests that being assigned an instructionally effective CT predicts teacher candidates being more instructionally effective themselves. Yet, existing qualitative literature and anecdotal evidence indicates that teacher candidates are often assigned to CTs who are not the most instructionally effective teachers in their schools or districts (Greenberg, Pomerance, & Walsh, 2011). There are many possible explanations for this. First, there is some evidence that teacher education programs often 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, in preparation). Recent research conducted on student teaching placements suggests that proximity to the program or the pre-service 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. District and school leaders are sometimes hesitant to select their most instructionally effective teachers to serve as CTs because this means rookie teachers will 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 (Mullman & Ronfeldt, in preparation; St. John, Goldhaber, Krieg, & Theobold, 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, in preparation). Similarly, St. John and colleagues (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 teacher education program 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 policymakers and EPP leaders. These concerns were subsequently confirmed during interviews – as part of a research study on the variation in clinical preparation – by EPP leaders responsible for designing and implementing clinical experiences across Tennessee (Mullman & Ronfeldt, in preparation). When asked about the basis for these concerns, some mentioned an unpublished report by the SAS Institute (2014) from a pilot study in Tennessee which 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).

The impact of mentoring on CTs’ instructional effectiveness

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

We are aware, though, of only one recent study that has directly tested the impact of supervising a candidate on teachers' instructional performance. In Washington, Dan Goldhaber and colleagues (2018) tested whether mentoring a candidate affected student achievement, and whether effects were heterogeneous 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. However, there were differential effects by quartile of prior performance; namely, for CTs in the lowest prior value-added quartile, mentoring a candidate had a large and negative effect on students' math achievement. The authors suggest that more effective CTs are able to "mitigate" the impact of letting an inexperienced pre-service teacher take over instruction in the classroom; this finding adds to the growing body of evidence that the most instructionally effective teachers should be the ones to serve as CTs. Though Goldhaber and colleagues (2018) found that effects of mentoring a candidate on concurrent achievement to be zero on average, they found modest, positive impacts on student math and reading performance in subsequent years. 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 and colleagues (2018), we are interested in the effect that mentoring a candidate has on teachers' evaluation metrics. The present study, though, also extends prior research by considering 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 more, 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. (2018), we consider heterogeneity of effects by prior performance level and we also consider heterogeneity by school type. 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 et al., under review). This is further supported by a UK review of mentoring programs for novice teachers,

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 policymakers in Tennessee and elsewhere – a central motivation for this study – is that serving as a CT can harm teachers' evaluation scores¹ 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 CT might impact their future performance, during post-service years.

Our Logic Model (see Figure 1) begins with the assumption that observation ratings and value-added to student achievement 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

¹ In this paper, 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.

classrooms beyond teaching quality, so are prone to manipulation, error, and bias. We know, for example, that observation ratings tend to be lower in classrooms with students who are lower-achieving, Black and Latinx, special education, 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, who the rater is (e.g., master rater versus principal), and content being taught (White, 2017; Whitehurst, Chingos, & Lindquist, 2014). Similarly, value-added to student achievement 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 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 impacted by mentoring a candidate, it is important then 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 (e.g., math, art). In other cases, CTs have multiple preps, sections or classrooms (e.g., secondary science teachers). In those cases, CTs may hand over lead teaching responsibilities in some classes but retain lead teaching in others. Given that school principals/leaders are likely to observe CTs during preps/classes that CTs are personally teaching, the decision about which classes/preps 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/subjects then they are likely to be observed and evaluated in less

challenging classrooms which could, effectively, boost evaluations. Thus, as we consider mechanisms by which mentoring may impact a teacher's own evaluations, we consider ways in which the quality of teaching may actually be impacted and 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.

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 rookie teachers tend to be less effective teachers which could lead to lower student achievement scores. Thus, 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.

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 only have one class/prep then presumably the evaluator will observe the CTs teaching the same population of students as those taught by their candidates. As a result of inexperienced candidates being responsible 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. 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. Perhaps the mostly likely mechanism is that mentoring a candidate effectively doubles the numbers of teachers in a classroom. This decreases the student-to-teacher ratio and means that students are likely to get 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 teacher education programs. 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, by placing a candidate in the 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 impacted. Hosting a candidate could be a form of professional development for CTs. By observing and providing feedback to a new teacher, or in engaging in planning and reflective conferences with them, CTs could 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. This learning could manifest in improved performance in subsequent years.

Of course, a CT's performance could also decline in post-service years. Hosting a candidate could be taxing, both personally or professionally, for a CT. This could lead to CT burn-out and exhaustion. Having to fulfil 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 lead to 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 view serving as a CT as a symbol of being an exceptional teacher 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; post-service evaluations may subsequently decline (when rater inflation is removed) even where the quality of teaching performance is consistent.

Research Questions

Drawing upon this Logic Model, in this paper 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:

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

RQ 2: Are the effects different for different groups of CTs?

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

Data

Data for this paper comes from a unique dataset of CTs collected by the Tennessee Department of Education. This dataset includes information from seventeen teacher education programs² 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

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.

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

Descriptive Statistics

Our analytic dataset 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 measures (Teacher Value-Added Assessment System, or TVAAS. See Voster, Guranio, and Woolridge, 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 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 to 2013-2014 school years³. Teachers in Tennessee are assessed multiple times per year using the Tennessee Educator Acceleration Model (TEAM), a rubric that includes four domains and multiple indicators within each domain. The four domains are instruction, environment, planning, and professionalism. Professionalism is only assessed one time, at the end of the school year. Multiple domains and

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

indicators are scored at the same time, and teachers receive scores on a scale from 1 (significantly below expectations) to 5 (significantly above expectations). For this paper, 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 to 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 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 (FRPL). 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 to 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., 2018; Matsko et al., 2018; Ronfeldt et al., 2018).

CTs Blocks

In order to conduct a more appropriate comparison of those who serve as CTs to those who do not, we constructed blocks of all eligible teachers for a student teaching placement in a given year. We identified teachers who served each year and then grouped them with all other teachers in their districts with the same teaching endorsement (e.g. secondary math, elementary, secondary ELA, etc.). This allowed us to create a hypothetical pool of all teachers who could have served for a

particular candidate.⁴ We merged Tennessee’s Personnel Information Reporting System (PIRS) and teacher assignment data onto our analytic sample and then compared the courses they taught that year and assigned them an endorsement. If for example, a seventh-grade social studies teacher in district *D* served as a CT, we create a block with that CT with all other secondary social studies teachers in that district, 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 to 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, on a scale of 1 to 5. Non-CTs in the same blocks have average observation ratings 0.19 lower than CTs. 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 to 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.

Methods

Research Question 1

We use a teacher fixed-effects model to investigate the effects of serving as a CT on evaluation metrics. This model allows us to estimate the within-teacher changes on years during

⁴ It is common for EPPs 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 (Maier & Youngs, 2009; Krieg et al., 2016). One reason for this is that candidates often have geographic and travel constraints.

which they serve as a CT as compared to the other years during which they did not mentor a teacher candidate.

Our preferred model is:

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

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, Krieg, and Theobald (2018) 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.⁵ We calculate effectiveness quartiles using a two-step approach. First, we estimate the teacher fixed-effect from this model:

⁵ Goldhaber et al. (2018) 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

$$Y_{it} = \tau_i + \pi_1 CT_{it} + \epsilon_i \quad (2)$$

where τ_i is the teacher fixed-effect for teacher i . It 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

$$Y_{it} = \beta_{0i} + \beta_1 CT_{it} \cdot Q_i + \delta Exp_{it} + \lambda_t + \epsilon_{it} \quad (3)$$

where $CT_i \cdot Q_i$ 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 are different for teachers at different points of the teacher performance continuum.

Heterogeneity by School Type. 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 (often in different subject or content areas) across the day. These differences also afford different opportunities for candidates placed at different school levels that could have implications for their CTs' evaluations. For example, elementary candidates likely have more opportunities to build relationships with students which could differentially benefit evaluations of elementary CTs over secondary CTs. Alternatively, secondary CTs may be able to strategically place candidates in courses or sections that could minimize potential costs and maximize potential benefits. Assuming school leaders will schedule observations during

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' estimates (Vosters, Guranio, & Woolridge, 2018).

classes/periods in which teachers are not mentoring a candidate, CTs could place candidates in their most difficult classes/periods. We felt an important first step was to investigate whether or not the effects of serving as a CT vary by school level.

Thus, we divide schools into four categories, elementary schools (grades K-5), middle schools (grades 6-8), high schools (grades 9-12), and other schools (other grade configurations, for example K-8). We estimate a model similar to (4) where we interact the CT indicator with indicators for school type. This allows us to test whether the effects of serving as a CT are concentrated in a specific instructional setting.

Research Question 3

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

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

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 one 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.⁶ 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 to 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.

⁶ 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.

First, we test whether our results are sensitive to the inclusion of teacher experience. Papay and Kraft (2015) argued that the 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 teachers' evaluation ratings to be related to the characteristics of their students (Campbell & Ronfeldt, 2018; Jiang & Sporte, 2016; Steinberg & Garrett, 2016).

Third, we submit our preferred model to 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 that 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 difference-in-differences and matched-sample model specifications to check whether our results are sensitive to model specification. In detail, we use the equation

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

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 model compares teachers who serve as CTs to 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 the group

of teachers that is never selected to serve as a CTs. 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 followed 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 were following parallel trends before serving as CT. In particular, when teachers served as CTs early in our observation window, we sometimes have no data on pre-trends, or we only have scores for a single year. Thus, we limit our analysis to the 2013 and 2014 CT cohorts since they have at least two years of pre-trend data, and we present an event study using this cohort (see Appendix Figure 1). While we observe parallel trends between CTs and non-CTs in terms of TVAAS, we observe that OR seem to increase the year prior to being selected to serve as a CT.

These results could hint at CTs being selected 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; we include difference-in-difference results in this paper as a robustness check to our preferred model. We recommend caution when interpreting the results from the 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 our difference-in-differences estimates have greater magnitude than our fixed-effects estimates. Two possible sources of bias can explain these results. First, if this model does not meet the parallel trend assumption, the estimates will be biased upwards. Second, another possible source of bias comes in the form of unobservable variables that

lead to an increase in observation ratings that is unrelated to serving as a CT. An example of this could be a teacher taking a leadership role (e.g., curriculum coordinator) at the school who then gets rewarded with mentoring a candidate. We can expect that taking on that role could lead to higher ORs and that is increase unrelated from 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 pre-trends 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 that have similar observable characteristics 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, post-bachelor'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 time of serving as a CT. We remove from these analyses two CTs who did not match with other teachers in the state on background characteristics.

Appendix Figure 2 reports the density distributions for the fuzzy matched variable pre- and post-matching. We note that the matching procedure was able to identify similar teachers across these three variables for all four outcomes of interest and that the common support assumption appears to have been met. We also note that we were not able to have quality matches on TVAAS mathematics scores two years prior to serving as a CT. This 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. 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

RQ1: 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 to 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 and 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.⁷ 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 dataset (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 to performance in years in which they do not, above we use the full sample of teachers – including those teachers in non-CT blocks – to estimate coefficients for teaching experience and for the intercept term. Thus, we wondered whether our estimates could be sensitive to our choice for the estimation 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,

⁷ As we discussed in the Methods 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. Our results 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 pre-trends are not parallel for observation ratings (see Methods and Appendix Figure 2). However, the results that we report in Appendix Tables 1 and 3 seem to suggest that that our estimates are well-specified.

subject areas, and years as CTs in our sample; and (3) only to teachers that 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 1). 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 descriptive statistics, described above, 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.⁸

Alternatively, this might indicate the presence of positive selection bias that is not fully accounted for in 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 similar observed characteristics to CTs but that were not picked to serve as CTs (see Methods for details). We report these estimates in Table 5, alongside the estimates for the teacher fixed effects and difference-in-difference models. Overall, we observe that the estimates for observation ratings have the same sign and magnitude across the different estimation models. It is notable that our matching algorithm matches on two years of prior evaluation trends, including on observation ratings. Thus, we are matching CTs with non-CTs that have similar patterns of returns to experience preceding the service years. Where differences in pre-trends could explain the positive effects on observation ratings in CT service years for our difference-in-difference specifications, they 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

⁸ 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.

difference-in-differences models, fail to account for self-selection bias. Said another way, teachers that were selected to be CTs could be different from other teachers 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. Failing to account for these unobserved differences could lead to estimates that are biased upwards.

RQ2: 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 types.

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 the CT indicators 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 to 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.⁹ There are other possible explanations as well, e.g., any contemporaneous professional development benefits of serving as a CT might also effect lower-performing CTs most.

⁹ Observation scores averages are 3.36 (s.d. = .36) for CTs in quartile 1, 3.73 (s.d. = .30) for teachers in quartile 2, 4.01 (s.d. = .28) for quartile 3, and 4.39 (s.d. = .38) for quartile 4.

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 4th quartile of effectiveness and possibly negative, but imprecisely estimated and nonsignificant, effects for teachers in the 1st quartile of effectiveness. Finding effects to be more negative for lower-performing teachers is consistent with what Goldhaber et al. (2018) found in their sample of teachers from Washington state, where they suggest more effective teachers are better able to buffer any potential negative effects of their mentees. However, 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 find 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., 2018). In this case, we might conflate 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 be at their peak performance then they will tend to regress to their mean performance in post-CT years; this could lead to estimates like the ones that we observe for observation ratings in Table 6.¹⁰ 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

¹⁰ 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 change 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.

effects of serving as a placebo CT for these 1,000 cohorts in order to us to test the extent to which our results are sensitive to regression to the mean. Appendix Table 3 shows the results from this Monte Carlo simulation. If regression to the mean were at play, then we would expect placebo CTs at the ends of the distribution in performance to have non-zero estimates. We find that that all the point estimates for the placebo CT sample are all close to zero and their 95% confidence intervals are centered at zero. Said in another way, we find that the placebo effect of serving as a CT on evaluation scores is centered around zero 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 Type. Table 7 displays the results for the effects of serving as a CT by school type. We find that the positive results on observation ratings are driven by CTs who teach in elementary and middle schools and that the evaluation scores of high school or other school teachers do not change when serving as a CT. Though we are not entirely sure why we observe these differences by school level, we discuss possible explanations below (see Discussion). 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.

RQ3: 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 we hypothesize in Figure 1, 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 post-service years. Specifically, we compare performance while serving as a CT and after serving as a CT to 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 estimate from the difference-in-differences model changes sign but remains non-significant. 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 4 column 3). This could suggest that CTs might have differential returns to experience on TVAAS scores (Atteberry et al., 2015, find differential returns to experience by quartile of performance). Specifically, 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 post-serving decline may be expected if non-CTs close the TVAAS gap during the post-CT period.

Similar to RQ1, we explore whether our results are sensitive to sample selection. Appendix Table 4 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 CT appears to move towards 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 towards zero.¹¹

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 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; moreover, program and district leaders charged with recruiting these teachers to serve can assure recruits that such fears are likely unwarranted.

¹¹ 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 one in our preferred model. The results from these models are consistent with the estimates from our preferred models (see Appendix Table 5), confirming a null effect on observation ratings for years following serving as a CT and a possible small and negative effect on TVAAS scores.

Since this paper 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. (2018). 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. (2018) 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 non-significant 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 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. One might be tempted to conclude, based on this latter finding, that program and school leaders should place candidates with lowest-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., 2018; Ronfeldt, Brockman & Campbell, 2018; Ronfeldt, Matsko et al., 2018). Second, results from our study and Goldhaber et al. (2018) indicate that student achievement gains among lowest-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 role as CT is somehow changing 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 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. (2018) and SASS (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 EPPs that kept and were willing to share these data, and only for years that

were included in their records. Thus, our coverage across EPPs and years was uneven and our results are not necessarily generalizable across the state. It is possible that the CTs for the particular EPPs 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, which we do not observe. 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 and colleagues (2018) 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 earlier, 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 – e.g., 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 planning 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.¹² 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 6).

It is true that secondary teachers also often have multiple preps (e.g., a science teacher may teach biology, advanced biology, and chemistry in the same term). Secondary teachers could also

¹² 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 teacher can prepare for planned observations, this may effectively boost evaluations for elementary CTs.

then assign their teacher candidates to the subjects or preps 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 or prep that is a match. Compared to 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 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, related, whether the increases reflect improvements in teaching quality or 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, work that is often unrecognized, unappreciated, 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., under review). 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.

Finally, our research extends prior work by testing whether or not serving as a CT has a longer-term effect on evaluation during post-serving years. A positive effect in post-serving years

could suggest that mentoring serves a professional development function. For observation ratings, we find post-service performance to decline back to pre-service levels. For TVAAS, we find post-service performance to actually be somewhat worse than preservice levels. However, when we constrain models only to individuals who ever served as CTs then the post-serving estimates are similar to pre-serving 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 potentially 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-serving 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. (2018), who find student achievement gains to increase significantly in post-serving 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 upon teacher-level TVAAS measures which do not adjust for student characteristics. If such differences in VAM construction were responsible for differences in effects on post-service 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-service 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 benefit them.

References

- Angrist, J. D., & Pischke, J.S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do First Impressions Matter? Predicting Early Career Teacher Effectiveness. *AERA Open*, 1(4), 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(4), 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). Retrieved from <https://search.proquest.com/docview/1612601875?accountid=14509>
- Campbell, S.L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*. Advanced online publication.
- 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*.
- Desimone, L.M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181-199.
- Gitomer, D. & Bell, C. (2013). Evaluating Teachers and Teaching. In Geisinger, K. F., Bracken, B. A., Carlson, J. F., Hansen, J. I. C., Kuncel, N. R., Reise, S. P., & Rodriguez, M. C. (Eds) *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education*. 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.

- Goldhaber, D., Krieg, J., & Theobald, R. (2018b). The costs of mentorship? Exploring student teaching placements and their impact on student achievement. CALDER Working Paper.
- Greenberg, J., Pomerance, L., & Walsh, K. (2011). Student Teaching in the United States. Retrieved from the National Council on Teacher Quality, www.nctq.org/edschoolreports/studentteaching.
- 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(4), 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(1), 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(3), 794-831.
- Jackson, C. K., & Bruegmann, E. (2009). Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics*, 1(4), 85-108.
- Jiang, J. Y., & Spörte, 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/publications/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, Washington: Bill and Melinda Gates Foundation. Retrieved from: http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.

- Kraft, M.A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88(4), 547-588.
- Krieg, J., Theobald, R., & Goldhaber, D. (2016). A foot in the door: Exploring the role of student teaching assignments in teachers' initial job placements. *Educational Evaluation and Policy Analysis*, 38(2), 364-388.
- Loeb, S., & Candelaria, C. (2012). How stable are value-added estimates across years, subjects, and student groups?. *The Carnegie Knowledge Network*.
- Matsko, K.K., Ronfeldt, M., & Greene Nolan, H. (under review). How different are they? Comparing preparation offered by traditional, alternative, and residency pathways.
- 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.
- Maier, A., & Youngs, P. (2009). Teacher preparation programs and teacher labor markets: How social capital may help explain teachers' career choices. *Journal of Teacher Education*, 60(4), 393-407.
- Makel, M.C., & Plucker, J.A. (2014). Facts are more important than novelty. *Educational Researcher*, 43(6), 304-316.
- Mullman, H., & Ronfeldt, M. (in preparation). The landscape of clinical preparation in Tennessee.
- 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(1), 3-26.
- Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education*, 66(4), 304-320.
- Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does Cooperating Teachers' Instructional Effectiveness Improve Preservice Teachers' Future Performance? *Educational Researcher*, 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. SAS Institute Inc.: Cary, NC, USA.
- 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(3-4).

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance:

What do teacher observation scores really measure?. *Educational Evaluation and Policy*

Analysis, 38(2), 293-317.

St. John, E., Goldhaber, D., & Krieg, J., Theobald, R. (2018). How the match gets made: Exploring

candidate placements across teacher education programs, districts, and schools. CALDER

Working Paper. Retrieved from

<https://caldercenter.org/sites/default/files/CALDER%20WP%20111018.pdf>

Vosters, K. N., Guranio, C. M., & Wooldridge, J. M. (2018). *Understanding and evaluating the SAS*

EVAAS univariate response model (URM) for measuring teacher effectiveness. UNC Charlotte Economics

Working Paper Series. Retrieved from [https://belkcollegeofbusiness.uncc.edu/economic-](https://belkcollegeofbusiness.uncc.edu/economic-working-papers/wp-content/uploads/sites/850/2018/06/wp2018-001.pdf)

[working-papers/wp-content/uploads/sites/850/2018/06/wp2018-001.pdf](https://belkcollegeofbusiness.uncc.edu/economic-working-papers/wp-content/uploads/sites/850/2018/06/wp2018-001.pdf)

White, M. (2017). *Generalizability of observation instruments*. (Unpublished doctoral dissertation).

University of Michigan, Ann Arbor, MI.

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). Evaluating teachers with classroom

observations. *Brown Center on Education Policy: Brookings Institute*.

Figure 1. Logic Model for Effects of Serving as a Cooperating Teacher on Evaluation Scores

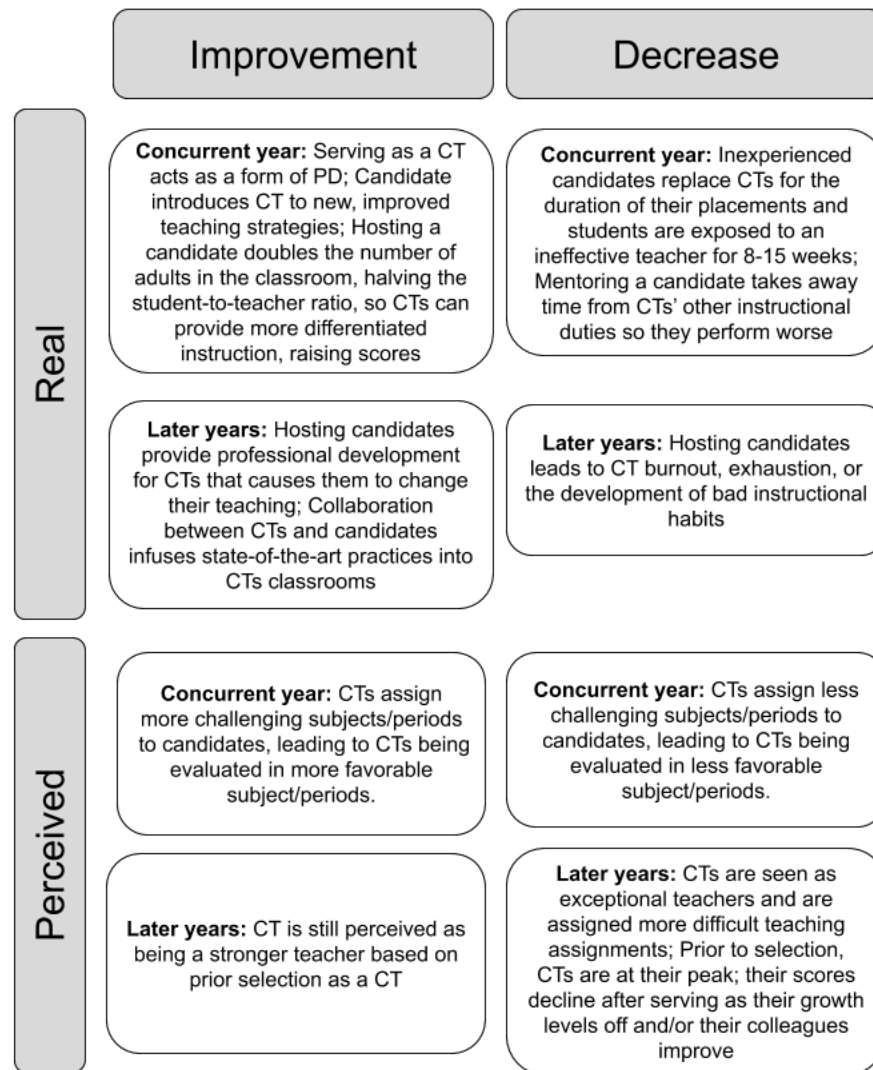


Table 1. Number of Teachers with Valid Evaluation Data

	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	<i>Cooperating Teacher Data Unavailable</i>			4,522
Observation Ratings	0	1,151	1,507	1,291				3,949
TVAAS Scores	417	472	983	786				2,658
Did Not Serve as Cooperating Teacher	21,291	73,349	70,195	72,525	<i>Cooperating Teacher Data Unavailable</i>			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. The Cooperating Teacher database is available for a subset of Educator Preparation 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 to Those of Cooperating Teachers

	All Teachers	Cooperating Teachers	Other Teachers	Diff	p-value
<i>Outcomes of Interest</i>					
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	***
<i>Teacher Covariates</i>					
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 Bachelors Degree	0.408	0.326	0.414	-0.088	***
Percent Masters 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	***
<i>School Assignment</i>					
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	***
<i>School Covariates</i>					
Percent White	0.678	0.759	0.673	0.086	***
Percent Black	0.216	0.140	0.221	-0.081	***
Percent Hispanic	0.075	.074	.075	-0.001	
Percent FRPL	0.587	0.567	0.589	-.0022	***
Percent Proficient	0.514	0.537	0.513	0.024	***
N	241,882	4,522	237,360		

Note. The Cooperating Teacher database is available for a subset of Educator Preparation 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. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 3. Descriptive Statistics by Block

	Blocks Without CTs	Blocks with CTs		
		All	Non CTs	CTs
<i>Outcomes of Interest</i>				
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
<i>Teacher Covariates</i>				
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 Bachelors Degree	0.401	0.419	0.429	0.330
Percent Masters 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
<i>School Assignment</i>				
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
<i>School Covariates</i>				
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. 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 pre-service teacher that year.

Table 4. Effects of Serving as a Cooperating Teacher on Evaluation Metrics

	Observation Ratings		TVAAS - All Subjects		TVAAS - Math		TVAAS - ELA	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Fixed Effects	Diff-in-Diff	Fixed Effects	Diff-in-Diff	Fixed Effects	Diff-in-Diff	Fixed Effects	Diff-in-Diff
Cooperating Teacher	0.040*** (0.007)	0.053*** (0.007)	0.008 (0.006)	0.014* (0.006)	-0.001 (0.012)	0.005 (0.013)	-0.003 (0.007)	0.005 (0.007)
Ever Cooperating Teacher		0.108*** (0.007)		0.040*** (0.006)		0.075*** (0.011)		0.028*** (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-Squared	0.771	0.292	0.689	0.125	0.702	0.194	0.619	0.131
Adjusted R-Squared	0.639	0.286	0.541	0.110	0.537	0.167	0.412	0.105

Note. 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 0-30 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 diff-in-diff models. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 5. Coefficient Sensitivity to Estimation Method

	(1)	(2)	(3)	(4)
	Observation Ratings	TVAAS All Subjects	TVAAS Math	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. This table reports the sensitivity of the cooperating teacher 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 (ATE) on teachers who serve as Cooperating teachers. 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. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 6. Heterogeneity by Quartile for RQ1

	(1)	(2)	(3)	(4)
	Observation Ratings	TVAAS All Subjects	TVAAS Math	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	233149	117034	52933	60082
R-Squared	0.748	0.662	0.671	0.585
Adjusted R-Squared	0.643	0.527	0.521	0.398

Note. 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 0-30 and as a pooled indicator for experience above 30 years. We drop singleton observations from models with teacher fixed effects. 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. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 7. Heterogeneity by School Type for RQ1

	(1)	(2)	(3)	(4)
	Observation Ratings	TVAAS All Subjects	TVAAS Math	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	170464	87080	38666	44243
R-Squared	0.771	0.689	0.700	0.618
Adjusted R-Squared	0.639	0.534	0.527	0.402

Note. 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 0-30 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 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. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

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*** (0.009)	0.049*** (0.009)	-0.002 (0.007)	-0.004 (0.008)	-0.010 (0.015)	-0.013 (0.015)	-0.011 (0.008)	0.001 (0.008)
After Cooperating Teacher	0.015 (0.009)	-0.005 (0.010)	-0.020* (0.008)	-0.030*** (0.009)	-0.016 (0.017)	-0.024 (0.016)	-0.017* (0.008)	-0.007 (0.008)
Ever Cooperating Teachers		0.112*** (0.009)		0.058*** (0.008)		0.091*** (0.014)		0.032*** (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 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	233149	242339	117034	127669	52933	61943	60082	69642
R-Squared	0.748	0.292	0.662	0.125	0.671	0.194	0.585	0.131
Adjusted R-Squared	0.643	0.286	0.527	0.111	0.521	0.167	0.398	0.105

Note. 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 0-30 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 diff-in-diff models. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

APPENDICES

Appendix Table 1. Robustness Checks of Teacher Fixed Effects Models on Different Subsamples of Teachers

	(1) Main Effects	(2) CT Blocks	(3) CT Schools	(4) CT Ever
<i>Panel A. Observation Ratings</i>				
Cooperating Teacher	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-Squared	0.771	0.784	0.816	0.741
Adjusted R-Squared	0.639	0.638	0.671	0.603
<i>Panel B. TVAAS – All Subjects</i>				
Cooperating Teacher	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	91726	52838	11817	7042
R-Squared	0.689	0.709	0.728	0.668
Adjusted R-Squared	0.541	0.537	0.522	0.517

Note. 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 0-30 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 diff-in-diff models. + p < 0.10 * p < 0.05 ** p < 0.01 *** p < 0.001

Appendix Table 2. Nearest Neighbor Matching Quality

	Mean		Variance	
	Raw	Matched Sample	Raw	Matched Sample
<i>Panel A. Observation Ratings</i>				
OR - Two Years Prior	-0.847	0.000	1.874	1.000
OR - One 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		
<i>Panel B. TVAAS</i>				
TVAAS - Two Years Prior	-0.097	0.012	0.503	1.135
TVAAS - One 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		
<i>Panel C. TVAAS Mathematics</i>				
TVAAS Math - Two Years Prior	-0.127	0.015	0.427	1.088
TVAAS Math - One 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		
<i>Panel D. TVAAS ELA</i>				
TVAAS ELA - Two Years Prior	-0.036	0.009	0.704	1.081
TVAAS ELA - One 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. This table reports the standardized difference between CTs and non-CTs on the variables we used to construct the nearest neighbor matched sample. Values close to zero for the matched sample means and close to 1 for the matched sample variance indicate that the matching procedure was able to identify a similar non-CT sample to the observed CT sample.

Appendix Table 3. Monte Carlo Simulation

	(1)		(2)		(3)		(4)	
	Observation Ratings		TVAAS - All Subjects		TVAAS - Math		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. This table reports the results of a Monte Carlo simulation that draws 1000 placebo CTs and calculates the placebo effect of serving as a 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. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Appendix Table 4. Robustness Checks of Growth Models on Different Subsamples of Teachers

	(1)	(2)	(3)	(4)
	Main Effects	CT Blocks	CT Schools	CT Ever
<i>Panel A. Observation Ratings</i>				
Cooperating Teacher	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	233149	81807	19880	13157
R-Squared	0.748	0.784	0.816	0.707
Adjusted R-Squared	0.643	0.638	0.671	0.599
<i>Panel B. TVAAS – All Subjects</i>				
Cooperating Teacher	-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	117034	52918	11817	8468
R-Squared	0.662	0.709	0.728	0.645
Adjusted R-Squared	0.527	0.537	0.522	0.515

Note. 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 0-30 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 cooperating teacher. Models (1) and (4) include singleton observations by teacher-year. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Appendix Table 5. Two-Stage Estimation for Cooperating Teacher Growth Trajectories

	Observation Ratings		TVAAS All Subjects		TVAAS Math		TVAAS ELA	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Teacher FE	Diff-in- Diff	Teacher FE	Diff-in-Diff	Teacher FE	Diff-in- Diff	Teacher FE	Diff-in- Diff
Cooperating Teacher	0.051	0.048	-0.002	-0.011	-0.015	-0.015	-0.012	-0.008
After Cooperating Teacher	0.023	-0.005	-0.021	-0.043	-0.025	-0.038	-0.020	-0.027
Ever Cooperating Teachers		0.112		0.066		0.098		0.043

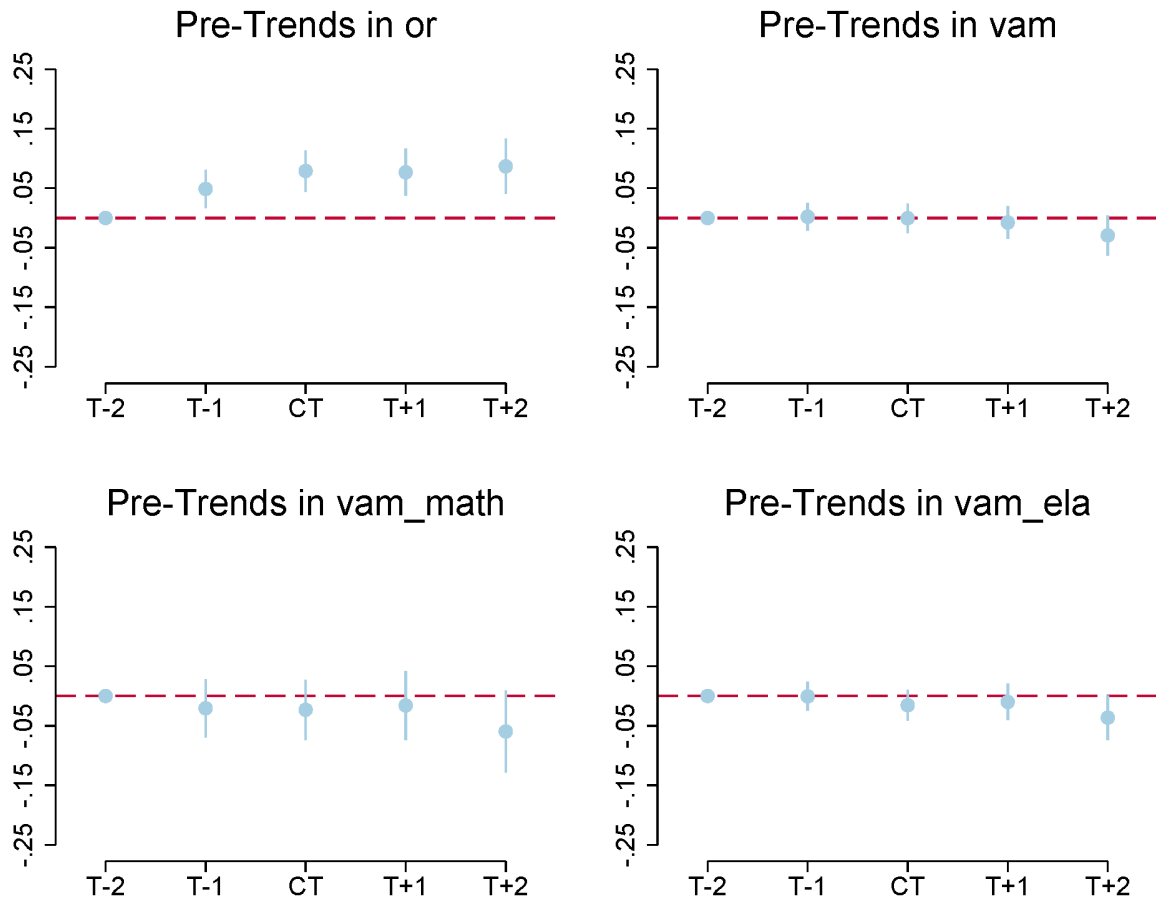
Note. Standard errors are not reported because they are not calculated for the two-stage Papay-Kraft estimation correction. Cooperating teacher is a time-varying indicator taking the value of 1 during the school year in which a teacher is reported as serving as a cooperating teacher. Experience is included as single indicators for years 0-30 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 diff-in-diff models.

Appendix Table 6. Heterogeneity by School Type and Quartile

	Observation Ratings				TVAAS - All Subjects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Elementary School	Middle School	High School	Other School	Elementary School	Middle School	High School	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		228417				112476		
R-Squared		0.749				0.663		
Adjusted R-Squared		0.644				0.523		

Note. 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 0-30 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 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. + $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Appendix Figure 1. Event Study of Serving as a CT on Outcomes of Interest



Appendix Figure 2. Density Distribution of Fuzzy Matched Covariates

