TN EDUCATION RESEARCH ALLIANCE

The Effects of More Frequent Observations on Student Achievement Scores

A Regression Discontinuity Design Using Evidence from Tennessee

Seth B. Hunter

Abstract

In the early 2010s, Tennessee adopted a new teacher evaluation system. Recent research finds Tennessee teacher effectiveness substantially and rapidly improved after this reform. However, there is little empirical research exploring which components of the reformed system might have contributed to this growth. Using longitudinal data, I apply a local regression discontinuity design to identify the effects of more frequent classroom observations, a cornerstone of Tennessee evaluation reform, on average student achievement scores. Much of the identifying variation is associated with an increase from one to two policy-assigned observations per year, potentially limiting the generalizability of results. However, most Tennessee teachers are assigned one or two observations by state policy, making this a margin of primary interest in the Tennessee context. Among teachers included in the research design, there is no evidence the receipt of an additional observation per year improved teacher effectiveness. Descriptive analyses suggest weak implementation of observational processes may explain the absence of positive effects. Implications are discussed.

Acknowledgements

Seth would like to thank Dale Ballou, the Tennessee Department of Education, and the Tennessee Education Research Alliance for valuable feedback.

WORKING PAPER
2019-04

1. Introduction

Since the mid-2000s, most state and large local education agencies have substantially reformed their teacher evaluation systems (Steinberg & Donaldson, 2016). Whereas evaluation systems under No Child Left Behind largely focused on school performance (Manna, 2011; Mehta, 2013), recently reformed evaluation systems focus on the teacher (Steinberg & Donaldson, 2016). Research produced over the 2000s suggested this new focus was warranted because analysts found teachers had a substantial impact on student achievement (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004) and teacher effectiveness varied substantially within schools (Aaronson et al., 2007; Rivkin et al., 2005). Soon after these findings became known, the federal Race to the Top competition incentivized education agencies to reform teacher evaluation systems to improve teacher effectiveness (US Department of Education, 2009).

Several state and local education agencies responded to these incentives (McGuinn, 2012), including Tennessee, one of the first recipients of a Race to the Top grant. Tennessee's reformed evaluation system further incorporated student outcomes into measures of teacher effectiveness, adopted a standards-based observation rubric (e.g. Danielson's Framework for Teaching), and increased the number of observations received by Tennessee teachers (Olson, 2018), among other reforms. Emerging evidence suggests Tennessee's reformed evaluation system has been successful on several fronts (Olson, 2018; Putman, Ross, & Walsh, 2018).

Notably, recent evidence from Tennessee suggests post-reform within-teacher returns to experience have been rapid, ongoing, and larger compared to the returns to experience observed in other settings (Papay & Laski, 2018). Evidence shows Tennessee teachers in the first five years on the job improved their effectiveness by approximately 0.08 and 0.18 standard deviations in reading and math, respectively. Between their fifth and fifteenth years, teacher effectiveness

improved an additional 0.02 and 0.05 standard deviations in reading and math. Compared to other settings, these improvements are large, especially the improvement in mathematics. Finally, relative to its previous system, the reformed Tennessee teacher evaluation system: maintained its impressive growth in mathematics, and increased the growth of within-teacher effectiveness in reading.

Given Tennessee's success in improving teacher effectiveness, practitioners and policymakers will want to know which components of the teacher evaluation system might have contributed to these successes. Indeed, there have been calls for such research (Jackson & Cowan, 2018). I address this call by identifying the effects of a cornerstone of Tennessee's reformed teacher evaluation system: the effects of receiving more classroom observations over a school year on average student achievement scores. Although there are other cornerstones supporting Tennessee's reformed evaluation system, there is reason to be concerned about the effects of the number of observations teacher received during a school year (i.e. "more frequent observations"). First, school administrators in similarly reformed systems report that observation system reforms are substantially time demanding (Kraft & Gilmour, 2016a; Neumerski et al., 2014) and quite burdensome (Rigby, 2015). These reports are unsurprising because the typical teacher in pre-reformed systems was observed once every few years, but is now expected to receive more than one observation each year (Steinberg & Donaldson, 2016). In Tennessee, the average teacher receives two observations each year. Moreover, previous research conducted outside Tennessee finds some administrators cope with observation-related demands by providing brief, low-quality observations and post-observation feedback conferences (Kraft & Gilmour, 2016a), potentially weakening the efficacy of more frequent observations. Second, local education agencies spend more on observations than any other component of reformed

teacher evaluation systems (Stecher et al., 2016). Combined with potentially weakened efficacy, the cost of observations may substantially lower the cost-effectiveness of these systems. The financial and administrative burdens associated with reformed observation systems underscore the importance of identifying the effects of more frequent classroom observations.

I identify the causal effects of more frequent formal observations on teacher effectiveness using longitudinal administrative data from more than 80 percent of Tennessee school districts. Treating variation in the number of observations received as exogenous is problematic (henceforth, "observations" refers to formal observations). For example, observers may observe less motivated teachers more often due to concerns about teacher effectiveness. To overcome this endogeneity problem, I exploit policy-imposed discontinuities in the assignment of classroom observations using a local regression discontinuity design. Because educators have no control over policy-assigned observations, variation in observations brought about by policy inducement is plausibly exogenous.

The identifying, policy-assigned discontinuity exists between the highest and next highest categories of Tennessee "overall" teacher effectiveness, with most teachers in these categories assigned one and two observations per year, respectively. Although these conditions suggest limited generalizability, this is not the case given the purpose of this analysis. The purpose of this analysis is to identify the effect of a cornerstone of Tennessee's reformed teacher evaluation system. Therefore, this analysis focuses on a component of Tennessee's system applying to a large share of Tennessee teachers. Because 70 percent of Tennessee teachers are assigned to the highest and next highest categories of overall effectiveness, the identification strategy uses data from a plurality of Tennessee teachers. Similarly, the marginal effect of receiving two observations instead of one per year is the margin applying to most Tennessee teachers.

Therefore, the effects identified at the margin between: the highest and next-highest categories of overall effectiveness, and one and two observations, are the effects of interest.

To preview my findings, the evidence implies that Tennessee teacher effectiveness did not improve because of more frequent observations. The receipt of more frequent observations left contemporaneous and longer-term average student reading scores relatively unchanged. Mathematics effects are estimated much more imprecisely: I am unable to rule out large negative contemporaneous effects, or small positive longer-term effects. However, in the preferred specification identifying longer-term effects on mathematics scores, estimates are either negative or near-zero. I conclude that there is no evidence average student math scores improved because of more frequent classroom observations. Descriptive analyses suggest the implementation of observational processes may be to blame for the absence of positive effects. Sizable minorities of teachers in the upper categories of overall teacher effectiveness report that they did not receive pre-observation conferences, and received post-observation feedback that was not useful for improving instruction.

The remainder of the paper is organized as follows. I discuss the study context, and methodology and data, in Sections 2 and 3, respectively. Threats to internal validity are discussed in Section 4. Section 5 describes findings and explores potential explanations for the results. Section 6 ends with conclusions, limitations, and implications.

2. Study Context

2.1 Tennessee Observation System Theory of Action

In the early 2010s, Tennessee made sweeping changes to its observation system, later named the Tennessee Educator Acceleration Model (TEAM). The TEAM theory of action

resembles those framing observation systems across the United States in that: 1) certified observers conduct observations using a standards-based rubric, 2) pre-observation conferences should precede "announced" observations, or observations the teacher knows about in advance, and 3) observers should share post-observation feedback in structured post-observation conferences following every formal observation. Although local education agencies could adopt alternative observation systems (Tennessee Board of Education, 2013), more than 80 percent adopted the TEAM observation system.

Observers must conduct observations using a rubric approved by TDOE (Tennessee Board of Education, 2013). Although local education agencies could use their own rubrics, over 80 percent use the state-default TEAM rubric (Tennessee Department of Education, 2016). The TEAM rubric (see Online Appendix 1) is a standards-based rubric (Alexander, 2016).

A pre-observation conference should precede announced observations (Tennessee Board of Education, 2013). Pre-observation conferences provide observers an opportunity to learn about the instructional goals of the upcoming lesson so they can anticipate teacher strengths and weaknesses (Alexander, 2016). During these conferences, teachers may request that the observer focus on specific student or teacher behaviors.

Post-observation feedback should be based on the observation rubric, provided during structured post-observation conferences, and should occur within one week of the observation (Tennessee Board of Education, 2013). Post-observation feedback is expected to improve teacher effectiveness as measured by student achievement, because research links the exemplary teacher/ student behaviors described in the TEAM rubric to higher student achievement (Daley & Kim, 2010). Teachers may be able to improve their effectiveness based on feedback alone, or

observers may direct teachers to suitable training opportunities (for example, workshops, teacher mentors, etc.).

Annual observer certification addresses the effective implementation of observational processes. Certification focuses on the accuracy and reliability of observation scores, and facilitation of pre- and post-observation conferences. Furthermore, certification is required to formally observe a teacher (Tennessee Board of Education, 2013).

2.2 Teacher Level of Effectiveness (LOE-cont)

After each school year, teachers receive a rating of their overall effectiveness, their discrete Level of Effectiveness (LOE). "Discrete LOE" is integer-scaled from one to five, but is based on an underlying continuous composite measure of teacher effectiveness combining teacher observation, "growth," and "achievement" scores. The growth score for all teachers in this analysis is based on their Tennessee Value-Added Assessment System (TVAAS) score. Achievement measures are determined by grade/ school/ district-wide student achievement (for example, ACT scores, high school graduation). I refer to the continuous composite measure determining discrete LOE as "LOE-cont." During the study period of 2012-13 through 2014-15, 50 percent of LOE-cont was determined by observation scores, 35 percent on teacher value-added, and the remainder was based on a grade-, school-, subject-, or district-wide student outcome (see Online Appendix 2 for details). LOE-cont ranges from 100 to 500 (Tennessee State Board of Education, 2013). TDOE assigns teachers whose LOE-cont is within [100, 200), [200, 275), [275,350), [350, 425), or [425, 500] to discrete LOE scores of 1, 2, 3, 4, or 5, respectively (Tennessee Department of Education, n.d.-b). Neither principals or teachers received teacher LOE-cont from any education agency during the study period.

2.3 Assignment of Observations

There are three broad factors affecting the number of observations teachers in the TEAM observation receive: certification status, prior-year LOE, and educator discretion. "Certification status" identifies whether a teacher has taught for less than four years ("Apprentice") or longer ("Professional"). TBOE assigns teachers with an LOE-cont greater than or equal to 425 one observation. The number of observations assigned to teachers below 425 depends on their certification status: Apprentice (Professional) teachers are assigned four (two) observations[i]. These represent the minimum number of observations a teacher should undergo, but districts/ schools can add to these minima. In general, TBOE expects each observation to take approximately 15 minutes (Tennessee Board of Education, 2013).

This analysis focuses on observations in the TEAM system given its widespread adoption and clear policies regarding the frequency of observations.

3. Methodology and Data

The main findings concern the contemporaneous effects of receiving more observations on teacher effectiveness. I also estimate longer-term effects because it is plausible it takes time for observations to affect teacher effectiveness. The outcomes of interest are average student achievement scores in math and reading. That is, the outcomes are teacher-year average student achievement scores.

3.1. Compliance with Treatment Assignment

If schools strictly adhered to TBOE guidelines, the number of observations teachers undergo would be a discontinuous function of their prior-year LOE-cont. I could then identify the effect of observations on teacher effectiveness using a one-stage regression discontinuity design. However, adherence is not perfect. Figure 1 is a binned scatterplot of the number of observations received against prior-year LOE-cont by certification status. The smoothed curves in Figure 1 are second-order polynomials of prior-year LOE-cont. This figure shows Apprentice (Professional) teachers with an LOE-cont greater than or equal to 425 tend to receive a total of two observations per year, or one (one-half) more observations than assigned by policy. Professional teachers with an LOE-cont below 425 tend to receive close to the policy-assigned two observations. However, Apprentice teachers with an LOE-cont below 425 tend to receive between three and one-half or three observations per year, or between one-half and one fewer observations than assigned by policy.

I characterize deviations between the number of observations received and policy-assigned number of observations as "non-compliance" with treatment, where treatment is the number of observations assigned to teachers by state policy. Tennessee policy assigns teachers one, two, or four observations (for a total of three potential treatments) depending on certification status and prior-year LOE-cont. Figure 1 illustrates that there is non-compliance for teachers above the 425-threshold, and Apprentice teachers below the threshold.

Because of non-compliance with TBOE guidelines (that is, non-compliance with treatment assignment), the number of observations received is plausibly endogenous. Teacher motivation is plausibly related to teacher effectiveness and the number of observations received. School administrators may observe less motivated teachers more often to closely monitor teacher

behaviors, negatively biasing estimated effects. Alternatively, school administrators may observe more motivated teachers more often because these teachers are receptive to feedback, introducing positive bias. Student behaviors may also influence how often a teacher is observed. For example, a teacher struggling with a difficult class may be observed more often.

3.2 Methodology

To estimate the effect of observations on teacher effectiveness, I employ 2SLS local regression discontinuity designs, relying on variation in prior-year LOE-cont surrounding the 425-threshold as an instrument to predict the number of observations received in year *t*. There are two instruments: whether an Apprentice teacher lies to the left or right of the 425-threshold, and an analogous instrument for Professional teachers. The relationship of interest is between the number of observations received over a school year and average student achievement:

$$(1)\ obs_{ijt} = \ddot{\beta}_0 + \boldsymbol{\theta} g(\cdot) + \ddot{\boldsymbol{A}} f(\cdot) + \ddot{\boldsymbol{B}} \boldsymbol{X}_{ijt} + \ddot{\boldsymbol{C}} \boldsymbol{S}_{jt} + \gamma_t + u_{ijt}, \qquad \left|LOE\text{-}cont_{ijt}\right| \leq w$$

$$(2)\ \ y_{ijt} = \beta_0 + \delta \widehat{obs}_{ijt} + \boldsymbol{A} f(\cdot) + \boldsymbol{B} \boldsymbol{X}_{ijt} + \boldsymbol{C} \boldsymbol{S}_{jt} + \gamma_t + e_{ijt}, \qquad \left|LOE\text{-}cont_{ijt}\right| \leq w$$

where y_{ijt} is the average grade-subject standardized math or reading achievement score for students taught by teacher *i* in school *j* in year *t*. obs_{ijt} is the number of observations received in year *t*. *f* is a second order polynomial of the running variable (prior-year LOE-cont) interacted with teacher certification status (Professional/ Apprentice) and *g* the vector of two instruments, such that the relationships between f and the outcomes are allowed to vary across the threshold. Variation in the instrumented number of observations received is based on the exogenous discontinuities in policy-assigned observations that occur when teachers cross the 425-threshold (see Figure 1). This procedure removes the plausibly endogenous variation brought about by educator discretion (see sections 2.3 and 3.1).

$\boldsymbol{X}_{ijt}$ is a vector of covariates including teacher race/ ethnicity, gender, years of teaching experience, certification status, level of education, and a fourth order polynomial of the prior-year average student achievement score of the same group of students. Stated differently, the prior-year polynomial is the student average achievement score measured in year $t - 1$ that is associated with the group of students taught by teacher i in year t. $\boldsymbol{X}_{ijt}$ also includes the proportions of the ith teacher's students according to: race/ ethnicity, free/ reduced price lunch status, ESL status, gender, and immigrant status. $\boldsymbol{S}_{jt}$ is a vector of school level measures controlling for the distribution of teacher effectiveness in school j in year t, including the mean, standard deviation, and skewness of prior-year LOE-cont. γ_t is a year fixed effect, and e_{ijt} and u_{ijt} are idiosyncratic error terms. The local average treatment effect of interest, δ, represents the effect of an additional observation per year on average student achievement scores. δ is estimated in bandwidths of 20, 30, and 40, which includes the Imbens-Kalyanaraman (2012) optimal bandwidth[ii]. Standard errors are clustered at the teacher level.

Although equations 1 and 2 use teachers in restricted bandwidths, these teachers represent a sizable share of Tennessee teachers. LOE4 and LOE5 teachers comprise approximately 70 percent of TEAM teachers. To the extent the estimates only generalize to teachers in the bandwidths, about 45 percent of TEAM teachers are in the largest bandwidth. Furthermore, the mean (standard deviation) prior-year discrete LOE of TEAM teachers is 3.96 (0.98).

Theoretically, summative observation scores represent an important intermediate outcome. To the extent more frequent observations change teacher effectiveness, those improvements should operate through improvements in teacher performance (i.e. observation scores). However, there is evidence of what I characterize as "observer bias" in observation

scores, introducing bias into the relationship between more frequent observations and observation scores (see Online Appendix 3). For this reason, I do not discuss the relationship between more frequent observations and observation scores any further. To be clear, observer bias cannot affect average student achievement scores, because students generate these scores. Observer bias only affects observer-generated observation scores.

3.2.1 Longer-Term Effects

Prior research implies more frequent observations should improve contemporaneous employee (teacher) productivity (Guerin, 1993; Kraft & Gilmour, 2016a; Murphy & Cleveland, 1995). However, the effects of more frequent observations may not materialize until subsequent academic years. Teachers may need time before they can incorporate post-observation suggestions into practice. Longer-term effects are especially plausible if educators use post-observation feedback to identify areas of weakness, then engage in training to develop teacher effectiveness.

I estimate longer-term effects by replacing $g(\cdot)$, $f(\cdot)$, and obs_{ijt} with $g_{t-1}(\cdot)$, $f_{t-1}(\cdot)$, and $obs_{ij,t-1}$, respectively. All controls and outcomes from equations 1 and 2 remain unchanged. Thus, longer term effects capture the impact of an additional observation per year on average student scores measured one year after treatment.

3.3 Data

To construct the predictors of interest, I draw on a rich set of administrative data from 2012-13 through 2014-15. Each record includes unique student, teacher, and school identifiers, allowing me to link students to teachers to schools. The data also include teacher and student

demographics (see Table 1 for a complete list of teacher and student demographics), teacher observation scores, LOE and LOE-cont, grades and subjects taught, and student achievement scores. Contemporaneous and prior-year measures of teacher effectiveness and performance, and student achievement, are included in each record. Records also include the percentage of subject-specific instructional time each teacher claims for each student taught. Teacher-year averages and proportions are weighted by these percentages. Test score data include scaled math and reading scores in grades three through eight, and high school end of course assessments in English I, II, and III, and algebra I and II. I standardize achievement scores by grade-subject.

Implementation and some robustness analyses use data from the Tennessee Educator Survey (TES). All Tennessee teachers receive the TES in late spring of each academic year. Response rates exceeded 50% during the study period.

Table 1 presents descriptive statistics for records used by equations 1 and 2 in a bandwidth of 40, the largest bandwidth. The typical (average) teacher is a white female, holds more than a BA/ BS degree, and has approximately 12 years of experience. Approximately 85 percent of records are associated with Professional teachers. Table 1 also shows roughly 40 percent of a math/ reading teacher's students are female, 12 percent are black, 65 percent are white, and six percent Hispanic. About 46 percent of a teacher's students have FRPL status, seven percent are ESL, and one percent have immigrant status. The typical teacher receives about two observations per year.

I examine if there are discontinuities in these characteristics at the 425-threshold by regressing each covariate used by equations 1 and 2 on the remaining covariates and two instruments. Although I can control for discontinuities in observable characteristics, the presence of several observable discontinuities would raise the concern that unobserved confounding

discontinuities also exist at the threshold. I find three discontinuities across 168 balance tests (see Table 2). When an Apprentice math teacher crosses from above to below the threshold, the proportion of her students that are FRPL increases by 0.07 to 0.12, which may negatively bias estimates. To the extent these imbalances introduce bias, I estimate some effects using Professional teachers only. Estimates produced by Professional-only and unrestricted samples are qualitatively similar: under no circumstance does the evidence imply that average student math scores improved because of more frequent observations.

4. Threats to Internal Validity

4.1 Manipulation of LOE-Cont

Because LOE-cont is determined in large part by scores from classroom observations conducted by school administrators, one might be concerned that administrators, wittingly or otherwise, manipulate scores such that teachers close to the 425-threshold fall to one or the other side. For example, an administrator might contrive to place a teacher above 425 because the administrator believes the teacher does not need additional observations. If manipulation is related to a teacher's subsequent effectiveness, as this example suggests, crossing the 425-threshold is not a valid instrument.

In fact, there is little reason to fear manipulation for this, or any other, reason. Administrators would need a keen prescience concerning teacher effectiveness to situate LOE scores just to one side of the 425-threshold because school administrators do not receive teacher measures of effectiveness until the completion of all observations. Thus, administrators wanting to manipulate LOE-cont, a measure they never received during the study period anyway (see section 2.2), would have to rely on historic discrete measures of teacher effectiveness to guess

current year scores. However, the correlation[iii] between prior-year growth and achievement measures (see section 2.2 and Online Appendix) from year t and $t-1$ is 0.50 and 0.37, respectively. These conditions suggest it is practically impossible for an observer to strategically place a teacher just to one side of the threshold. Moreover, there are no discontinuities in prior-year observation scores at the 425-threshold (see Table 2). Despite these conditions, I test for manipulation using the robust-bias correction approaches developed by Cattaneo, Jannson, and Ma (2016). There is no evidence of manipulation.

Conventional tests of manipulation compare the probability density function (PDF) of the running variable as it approaches a cut score (threshold) from the left, to the PDF of the running variable as it approaches the cut score (threshold) from the right. A relatively large difference between these PDF estimates is evidence of manipulation. However, this type of discontinuity in LOE-cont is expected because it is approximately continuous, invalidating conventional tests. LOE-cont scores are a weighted average determined by three components, and two components (achievement and growth scores) are integer-scaled from one through five. Five is the least common multiple of weights applied to these two integer variables. Thus, any LOE deviating from a multiple of five only does so due to observational ratings because this LOE component is composed of non-integer numbers. Stated differently, the distribution of observational ratings entirely determines the approximate continuity of LOE-cont.

Figure 2 is a histogram of LOE-cont. Figure 3 is the distribution of these same scores transformed via modulus five (LOEmod5). Figure 3 shows concentrations of the PDF at multiples of five, as expected. Considering the approximately continuous properties of LOE-cont, I remove the two integer components from LOE-cont before testing for manipulation.

I assume observers were prescient and knew teacher achievement and growth scores in advance, an unrealistic assumption. If there is no evidence of manipulation under this assumption, manipulation of LOE-cont under more realistic conditions is even more implausible. I use robust-bias correction approaches to test for manipulation (Cattaneo et al., 2016). The null hypothesis of no manipulation cannot be rejected ($p \sim 0.14$).

4.2 Validity of Instruments

The instruments are invalid if crossing the 425-threshold systematically affects average student achievement outside observational processes. Prior research and several conversations with the Tennessee Department of Education raised three potentially threatening mechanisms that may exist at the threshold: alternative policy treatments, systematic re-assignment of teachers to new positions, or teacher motivational responses to LOE assignment.

Crossing the 425-threshold does not trigger any statewide policies aside from the treatments of interest. Tenure is partially determined by crossing the 275-threshold, and there is no Tennessee policy concerning teacher bonuses or strategic compensation for crossing from the penultimate to highest category of overall teacher effectiveness. However, Tennessee districts could adopt their own strategic compensation policies during the study period, and some did. Yet, district-adopted bonus policies are only concerning if there is evidence they affected student achievement in mathematics or reading. In Tennessee, there is no such evidence (Ballou et al., 2017).

Despite the absence of threatening alternative policies at the 425-threshold, crossing the threshold may lead district or school leaders to assign teachers to new positions or schools[iv]. Switching grades, subjects, or schools may introduce new instructional challenges, negatively

affecting average student achievement. Alternatively, administrators may change teacher assignment because the new assignment is a better fit for the teacher, potentially introducing positive bias. Because of these concerns, I identify the effects of crossing the 425-threshold on three binary measures of teacher job assignments. Recall, educators never received LOE-cont, and received discrete LOE scores in early- or mid-fall, after initial teacher assignments. However, administrators could still respond to LOE scores received in the fall of year *t* by changing teacher assignments before the end of year *t*.

Robustness tests reveal that crossing the 425-threshold does not affect teacher assignment to a different school, or assignment from an untested subject in year $t - 1$ to a tested subject in year *t.* However, there is evidence when Apprentice reading teachers cross from below to above the 425-threshold they are more likely to teach a different grade/ subject (e.g. 3rd to 4th grade reading, high school English I to English II) in year *t*. To the extent this introduces bias, I estimate some effects using only Professional teachers. Again, estimates produced by unrestricted and Professional-only samples are qualitatively similar: there is no evidence average student reading scores improved because of more frequent observations.

The third broad threat to the validity of the instruments concerns teacher motivation. There are two plausible, potentially threatening motivational effects: an *impetus to improve* and *demoralization*. Teachers just below the threshold may face an impetus to improve, independent of the observation process. This impetus may exist due to socio-professional pressure (for example, teachers in higher discrete LOE may have more prestige) and would induce upward bias. An impetus to improve could originate from the teacher, her peers, or her administrators. At the same time, assignment to a lower discrete LOE may induce demoralization, regardless of the

observation process. Demoralization may lead teachers to apply less effort, which could lower student achievement, resulting in negative bias.

If either psychological threat exists, evidence to that effect should appear in teachers' responses to TES (survey) items concerning their improvement efforts. The TES asks about the following teacher improvement efforts: professional development[v], peer collaboration, preparation for classroom observations, instructional improvement, and extent to which teachers exerted effort on various activities (for example, lesson prep, reflecting on teaching). Thus, evidence that teachers below the 425-threshold spent more *or* less time in professional development, peer collaboration, etc. could threaten the validity of the instruments. Robustness tests find no systematic evidence of either motivational threat.

After the presentation of main findings in Section 5, I discuss more tests and results concerning the validity of the instruments, which also corroborate the internal validity of the research design.

4.2.1 Robustness Tests Concerning Job Assignment and Motivational Effects

Three variables capture whether a teacher switched: (1) grade/ subject, (2) from an untested to a tested grade/ subject, or (3) schools. The first captures whether the teacher switched to a new grade/ subject in year t relative to their grade/ subject in year $t - 1$. The second represents whether the teacher taught an untested subject in year $t - 1$. Although all teachers in the analytical sample taught tested subjects in year t, they did not necessarily teach a tested subject in the prior-year. The third assignment variable captures whether a teacher switched schools from year $t - 1$ to year t.

I examine discontinuities at the 425-threshold by regressing each assignment variable on the control and instrumental variables from equations 1 and 2 (see Table 3). There is no evidence of switching from an untested to tested subject at the 425-threshold. Nor is there evidence of school switching. Although neither instrument individually predicts changes in these two outcomes, the instruments could jointly predict switches in tested status or school switching, which may still introduce bias. However, there is no evidence the instruments individually or jointly predict these two outcomes, which could introduce bias (see Table 3). Yet, there is evidence that when an Apprentice reading teacher crosses from above to below the threshold they are less likely to switch grades/ subjects, but no corresponding discontinuity among math teachers or Professional reading teachers. To the extent grade/ subject switching among Apprentice reading teachers introduces bias, I restrict some samples to Professional reading teachers only. This restriction does not alter my conclusions.

I test for the presence of potentially biasing motivational effects using TES items. The TES asked teachers to report the number of hours spent in professional development. A second set of items asked teachers to list the number of times they collaborated with other teachers for various purposes (for example, improve instruction). Two more items asked about the amount of time teachers spent: 1) preparing for classroom observations, and 2) trying to improve their instruction. A final set of items asked teachers about the extent to which they exerted more time or effort on various activities. Online Appendix 4 contains these items, the original scales, descriptions of the transformations of these items into the five survey outcomes of interest, and descriptive statistics.

Robustness tests regress[vi] survey outcomes on the instruments and covariates from equations 1 and 2 in a single-stage equation. An instrument may be invalid if it predicts the

survey outcomes. In results not shown, neither instrument individually predicts any survey outcome. However, this does not mean the instruments do not jointly predict survey outcomes, which could also introduce bias. F-tests for joint significance find the instruments do not jointly predict any survey outcome (see Table 4).

Results in this section may be sensitive to the operationalization of survey outcomes. Sensitivity analyses using different operationalizations produce qualitatively similar results (see Online Appendix 5).

5. Findings

5.1 Main Findings

The bottom portion of the top panel in Table 5 shows the instruments strongly predict the number of observations received. The effects of an additional observation per year on contemporaneous average student math scores are negative, and the effects on average reading scores hover near zero. However, none of the relationships are statistically significant. To the extent identification assumptions are not met for Apprentice teachers, the bottom panel of Table 5 displays estimates using only Professional teachers. Professional-only estimates resemble the original estimates. All effects on math scores are negative, all effects on reading scores are near-zero, and none of the estimates are statistically significant at conventional levels.

Table 6 presents the longer-term effects of observations. The top panel displays results using Professional and Apprentice teachers. Longer-term effects on math teachers are insignificant and fluctuate around zero, ranging from 0.05 to -0.03. The corresponding effects on reading teachers are near-zero and insignificant. The bottom panel of Table 6 restricts the sample to Professional teachers for aforementioned reasons. Each of the longer-term effects among only

Professional teachers are more negative than the corresponding longer-term effect in the unrestricted sample, but none of the Professional-only effects are significant.

5.2 Robustness Tests: Demoralization and Leaving Teaching

There is no evidence that the receipt of more frequent observations substantially improved average student scores in math or reading. An impetus to improve, which would positively bias estimates, cannot threaten this conclusion. However, the negative bias associated with demoralization could partially explain these results. Although robustness tests in Section 4.2.1 found no evidence of motivational effects on survey outcomes, I conduct additional tests for demoralization considering the threat it poses.

If learning about assignment to discrete LOE4 (LOE-cont less than 425) in early- or mid-fall of year *t* induces substantial demoralization, this could lead teachers to leave teaching after year *t* (i.e. the teacher is not in the *t* + 1 administrative dataset). I regress whether a teacher leaves teaching after year *t* on the right-hand side variables in equation 1. Although point estimates exhibit expected patterns (see Table 8), all estimates are statistically insignificant. Thus, the collection of robustness tests effectively rules out threats posed by demoralization.

5.3 Robustness Tests for Generic Effects Related to LOE Assignment

All threats to the internal validity of the instruments share a common feature: assignment to a lower discrete LOE causes a change in teacher effectiveness independent of the observation process. If threatening effects exist at the 425-threshold, they should exist at other thresholds without discontinuities in the number of policy-assigned observations. To the extent assignment from a higher to lower discrete LOE induces a generic response affecting average student scores,

evidence of this response should exist when crossing from an LOE4 to LOE3 (LOE3 to LOE2), where there are no policy-assigned discontinuities in observations. Thus, I test for the presence of a generic LOE-assignment effect at each of the discrete LOE2/ 3 (LOE-cont 275) and LOE3/ 4 (LOE-cont 350) thresholds using local regression discontinuities, regressing average student achievement scores on the controls in equation 1 and a binary variable indicating whether a teacher is below or above the LOE-cont 275- (350-) threshold.

There is no evidence of a generic response to crossing discrete LOE threshold (see Table 9): all estimates are insignificant and most are near-zero. Furthermore, these new estimates provide additional evidence that the receipt of an additional observation per year did not benefit student achievement. The estimated effects on average student math scores at the 275- and 350-thresholds, where there are no additional policy-assigned observations, are the same or more positive than the corresponding estimates at the 425-threshold, where there are discontinuities in policy-assigned observations. This implies the effect of receiving an additional observation per year is near-zero or negative. A similar pattern exists among reading teachers. Crossing a discrete LOE threshold where there are no discontinuities in observations has a near-zero effect on average student achievement (see Table 9). However, crossing the 425-threshold, where teachers are assigned more observations, produces almost identical results, implying the effect of observations on average student achievement is near-zero.

5.4 Implementation of the Observational Processes

The absence of positive effects may exist because of weakly implemented treatments (i.e. observational processes). Recall, TDOE expects each observation to last approximately 15 minutes (see section 2.3), which may not be long enough to generate impactful post-observation

feedback. Weakly implemented pre- and post-observation conferences may also explain the absence of positive effects. To better understand why more frequent observations did not improve average student achievement, I present descriptive findings regarding implementation. Although TDOE did not collect data concerning comprehensive implementation throughout the study period, I use available data to explore if the findings are explained by the: 1) implementation of pre- or post-conferences, or 2) timing of additional observations.

The TEAM theory of action asserts pre- and post-observation conferences play an important role in the observation system (see Section 2.1). In 2013 and 2014, the Tennessee Educator Survey included items about the implementation of both conferences. To determine if implementation of pre- and/ or post-conferences might explain the mostly null findings, I add these survey items to equations 1 and 2. Doing so reduces the original analytical samples between 70 and 98 percent, drastically weakening power and casting doubt on the generalizability of results. The reductions in sample size occur because: 1) TDOE nests several survey items in modules administered to subsamples of Tennessee teachers, and 2) none of the items were administered across all three years of the study period. Due to the limitations of these regressions, I only discuss descriptive analyses about the implementation of pre- and post-conferences. Descriptive analyses are restricted to responses from discrete LOE4/ 5 teachers, the discrete LOE straddling the 425-threshold.

To the extent survey responses generalize to LOE4/ 5 teachers, descriptive analyses suggest most teachers assigned to higher categories of overall effectiveness receive pre- and post-observation conferences, but sizable minorities: 1) do not receive pre-conferences, and 2) receive feedback that is difficult to use for instructional improvement (see Table 10). Nearly one in five respondents (n = 4,767) report spending no time in pre-observation conferences. The

average amount of time[vii] respondents report spending in pre-conferences per observation ranges from 11 to 43 minutes. A similar item asks teachers about time spent "receiving/ reviewing" post-observation feedback. Nearly every respondent (98 percent) reports receiving/ reviewing post-observation feedback, and the average amount of time spent receiving/ reviewing feedback ranges from 13 to 52 minutes per observation. Furthermore, when observers discuss post-observation feedback with teachers, approximately 90 percent of respondents (n = 18,381) agree that their observer uses the TEAM rubric as the basis of the discussion. This is an important finding because the theory of action undergirding new observation systems asserts feedback based on standards-based rubrics (i.e. the TEAM rubric) will improve teacher productivity (see section 2.1).

Despite receiving/ reviewing post-observation feedback, survey responses suggest many teachers assigned to higher categories of overall effectiveness do not find the feedback helpful. Approximately 40 percent (n = 3,717) of respondents agree that it is difficult to use post-observation feedback to improve their practice. This difficulty may impede teacher take-up of observer recommendations based on post-observation feedback. Indeed, 28 percent (n = 2,981) of teacher respondents report not taking steps to improve their practice in their weakest area of teaching, as identified by their observer.

I also explore the extent to which the timing of observations explains the absence of positive effects. Teachers receiving more observations must have different observation schedules than teachers receiving less. Observers may cope with the demands of policy-assigned observations by conducting multiple observations of the same teacher in bursts. For example, an observer may observe a teacher assigned more frequent observations twice in one week,

undercutting the effects of the first observation on teacher effectiveness. Such bunching may account for the null findings.

I explore this potential explanation by finding the fraction of a teacher's total observations received within each of six two-month windows (for example, August-September), then include these fractions as right-hand side variables. Each record in Tennessee administrative data includes an "observation date," which I use to calculate the fractions. If the timing of observations explains the results, the new estimates would be significantly more positive than the main findings. However, the new results are statistically indistinguishable from the original (see Online Appendix 8). Furthermore, results are insensitive to the pairing of months used to construct the two-month windows (Aug-Sept versus Sept-Oct), and to the use of one-month windows.

6. Conclusions and Implications

There is emerging evidence that Tennessee's reformed teacher evaluation system has been successful on several fronts (Olson, 2018; Putman et al., 2018). Indeed, recent evidence shows teacher effectiveness in the reformed system improved rapidly and substantially (Papay & Laski, 2018). Considering that several reformed teacher evaluation systems in other contexts have not produced such positive results (Walsh, Joseph, Lakis, & Lubell, 2017), Tennessee's successes are laudable. Therefore, it is important to identify the effects of individual components of Tennessee's teacher evaluation system, so that other education agencies may replicate Tennessee's success by eschewing ineffective components and adopting effective components.

This paper examined a cornerstone of Tennessee's reformed system: an increase in the frequency of teacher observations. Tennessee, and several other state and large local education

agencies, presumed more frequent observations would more rapidly improve teacher performance and effectiveness by providing teachers with more frequent feedback for performance improvement (Steinberg & Donaldson, 2016). However, the evidence suggests that the receipt of more frequent, formal classroom observations by a large share of Tennessee teachers did not improve average student achievement in math or reading. Contemporaneous effects on average student math and reading scores, and longer-term effects on reading scores, are negative or near-zero, and all are statistically insignificant. One bandwidth produced a statistically insignificant, moderately sized, positive longer-term effect on math scores. However, robustness tests suggested the data in that bandwidth may not meet identification assumptions. When using data meeting identification assumptions, all longer-term effects on math scores are near-zero or negative. Due to the absence of clear positive effects, I interpret the evidence to mean that the receipt of more frequent observations did not improve average student math achievement. Descriptive evidence suggests the absence of positive effects may exist due to weak implementation of observational processes. That is, large shares of respondents to a statewide survey reported: not receiving any pre-observation conferences, and not using post-observation feedback to improve instruction.

6.1 Limitations

There are four potential limitations regarding generalizability, and one limitation regarding explanatory mechanisms. One may be concerned that because estimates are based on teachers surrounding the LOE-cont 425-threshold (i.e. discrete LOE4 or 5 teachers), the estimates will not generalize to other teachers. But, given the purpose of this analysis, this is not a concern. The purpose of this analysis is to explore the effectiveness of an important component

of the Tennessee evaluation system. Thus, I explored a component that applied to a large share of Tennessee teachers. A plurality of Tennessee teachers are assigned to these higher categories of overall effectiveness: discrete LOE4 and LOE5 teachers comprise nearly 70 percent of the population of TEAM teachers. Despite the representativeness of discrete LOE4 and LOE5 teachers, it is possible results only apply to teachers in restricted bandwidths. In this case, approximately 45 percent of TEAM teachers fall in the largest bandwidth used by local regressions. In short, the negatively skewed distribution of LOE implies the findings are not limited to a small share of teachers.

The source of identifying variation may represent a limitation. Observations occur for two broad reasons: policy-assignment and/ or educator discretion. All estimated relationships are local average treatment effects, identifying the effects of policy-assigned observations. Although school administrators may conduct policy-assigned observations merely to comply with state policy, it is plausible observations brought about by educator discretion would not occur unless they were a productive use of time. Therefore, discretionary observations may affect teacher effectiveness differently than policy-assigned observations.

The third generalizability-related limitation concerns outcomes. In this analysis, measures of teacher effectiveness are restricted to test-based outcomes, but teachers are responsible for more than improving student achievement. Moreover, observations may affect other outcomes because the TEAM rubric describes academic and non-academic behaviors. Future work may address non-academic outcomes of interest including student: disciplinary infractions, attendance, and course-taking.

Fourth, it is important to remember that no teacher in this analysis received zero observations. Therefore, estimates do not warrant conclusions about whether teachers should or

should not receive formal observations. Prior work already implies the receipt of at least one observation in new observation systems is better than no observations (Steinberg & Sartain, 2015; Taylor & Tyler, 2012).

Finally, a full investigation of implementation would clarify why more frequent observations do not improve teacher effectiveness among teachers assigned to higher categories of overall effectiveness. Future research may explore the importance of better defined observational processes and supports.

6.2 Implications

The evidence in this paper implies that Tennessee teacher effectiveness did not improve because of more frequent observations. These findings imply two courses of action. First, Tennessee and other education agencies may be able to improve the efficacy of more frequent observations if policymakers invested additional resources in implementation. However, given what is known about recently reformed observation systems, this seems unwise, at least in the short-term. Reformed observation systems are already the costliest component of teacher evaluation (Stecher et al., 2016); therefore, policymakers may resist further investment in these systems. Additionally, research implies school administrators cope with existing observational burdens by engaging in satisficing behaviors to comply with policy mandates (Halverson, Kelley, & Kimball, 2004; Kimball, 2003; Sartain et al., 2011), similar to those reported in the descriptive analyses. There is little reason to believe school administrators would react differently to additional, observation-related burdens brought along with additional training.

Although there are some challenges associated with the first implication, policymakers, teachers, and administrators may welcome the second implication. This analysis implies

policymakers wishing to replicate Tennessee's growth in teacher effectiveness can do so by assigning relatively effective teachers only one observation per year. Teachers assigned to higher categories of overall effectiveness may not learn new and useful information about their practices through additional observations as implemented by the typical observer. Indeed, it may be difficult for the typical observer (i.e. school administrator), who cannot possess expertise in all observed content areas, to even recognize when or if relatively effective teachers commit instructional mistakes. Such a scenario may explain why: almost 45 percent of Tennessee teacher respondents assigned to the higher categories of overall effectiveness found it difficult to use post-observation feedback to improve their practice, and nearly 30 percent of these teachers reported not taking any steps to address the weakest area of instruction identified by their observer.

In the case of Tennessee, this means teachers in the penultimate category of overall teacher effectiveness do not need two or four observations per year. These reductions would allow administrators to reallocate their time. In the Tennessee context, assigning all higher performing teachers (that is, those with a prior-year discrete LOE3 or above) one observation would reduce the total number of policy-assigned observations at the typical Tennessee school by thirteen observations per year. Administrators could use the time they would have spent observing these relatively effective teachers observing relatively ineffective teachers. If less effective teachers struggle with generic teaching practices, feedback from the typical (that is, generalist) observer may still benefit these teachers. Alternatively, administrators could use the time gained by not observing more effective teachers on other tasks.

Presumably, policymakers wishing to replicate Tennessee' growth of teacher effectiveness will also welcome the second course of action. It seems education research rarely

implies that policymakers and practitioners can do less without causing harm. Yet, evidence from Tennessee implies policy can assign relatively effective teachers just one observation per year without decreasing teacher effectiveness.

Tables and Figures

Table 1

Sample Descriptive Statistics

	Math Mean	Math SD	Reading Mean	Reading SD
Obs Received	1.75	(0.98)	1.87	(0.96)
TLM Math Scores	-0.01	(0.61)	.	.
TLM RLA Scores	.	.	-0.01	(0.55)
Female	0.84		0.90	
BA+	0.58		0.61	
Years Experience	11.85	(8.81)	12.53	(9.15)
Non-White	0.05		0.05	
% Sample from Professional Teachers	86.3	.	88.3	.
Proportion of Students Taught with Characteristics				
Female	0.43	(0.15)	0.38	(0.17)
Black	0.12	(0.17)	0.11	(0.16)
White	0.69	(0.27)	0.63	(0.28)
Asian	0.02	(0.04)	0.02	(0.04)
Hispanic	0.06	(0.09)	0.06	(0.09)
FRPL	0.49	(0.24)	0.44	(0.24)
ESL	0.07	(0.11)	0.06	(0.12)
Immigrant	0.01	(0.05)	0.01	(0.05)

Note: Descriptive statistics use data from the bandwidth 40 analytical sample. Standard deviations in parentheses. BA+ indicates whether teacher earned more than a BA/ BS degree. Non-White indicates whether teacher is black instead of white. Proportions represent the proportion of students taught with a given characteristic.

Table 2

Covariate Balance Tests

	Math Teachers			Reading Teachers		
Covariate	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
Prior-Yr Avg Student Achievement Score: App	0.07	0.14	0.07	0.05	0.09	0.03
	[0.200]	[0.125]	[0.101]	[0.167]	[0.112]	[0.091]
Prior-Yr Avg Student Achievement Score: Prof	-0.01	-0.02	-0.02	0.01	0.01	0.03
	[0.042]	[0.035]	[0.030]	[0.034]	[0.028]	[0.024]
Experience: App	0.51	0.32	0.21	1.75	1.30	0.47
	[1.101]	[0.816]	[0.664]	[1.144]	[0.811]	[0.619]
Experience: Prof	-0.53	0.55	0.86	-1.06	-0.00	0.25
	[0.914]	[0.747]	[0.655]	[0.937]	[0.758]	[0.647]
Female: App	0.08	0.08	0.02	> -0.01	0.02	0.04
	[0.090]	[0.076]	[0.066]	[0.086]	[0.072]	[0.061]
Female: Prof	-0.03	-0.03	-0.01	-0.04	-0.04	-0.02
	[0.040]	[0.032]	[0.028]	[0.029]	[0.023]	[0.020]
BA+: App	-0.17	-0.10	-0.11	-0.03	-0.01	< 0.01
	[0.113]	[0.093]	[0.080]	[0.129]	[0.102]	[0.085]
BA+: Prof	0.07	0.03	0.01	0.06	< 0.01	-0.01
	[0.051]	[0.042]	[0.037]	[0.047]	[0.039]	[0.033]
Non-White: App	-0.07	-0.06	-0.02	-0.11	-0.08	-0.06
	[0.049]	[0.040]	[0.034]	[0.057]	[0.046]	[0.039]
Non-White: Prof	-0.04	-0.02	-0.03	-0.02	< 0.01	> -0.01
	[0.024]	[0.020]	[0.016]	[0.021]	[0.017]	[0.015]
Prop of Students Taught with Characteristics						
Female: App	-0.01	-0.01	> -0.01	-0.03	0.01	0.01
	[0.034]	[0.026]	[0.021]	[0.035]	[0.026]	[0.022]
Female: Prof	-0.01	< 0.01	< 0.01	< 0.01	< 0.01	> -0.01
	[0.012]	[0.010]	[0.008]	[0.010]	[0.008]	[0.007]

Black: App	< 0.01	-0.01	-0.01	0.02	< 0.01	< 0.01
	[0.026]	[0.021]	[0.018]	[0.033]	[0.025]	[0.021]
Black: Prof	< 0.01	> -0.01	> -0.01	-0.01	-0.01	> -0.01
	[0.011]	[0.009]	[0.007]	[0.009]	[0.008]	[0.007]
White: App	> -0.01	> -0.01	< 0.01	-0.04	-0.04	-0.02
	[0.035]	[0.027]	[0.023]	[0.041]	[0.031]	[0.026]
White: Prof	> -0.01	-0.02	-0.01	-0.02	-0.01	-0.01
	[0.014]	[0.012]	[0.010]	[0.013]	[0.011]	[0.009]
Asian: App	< 0.01	< 0.01	< 0.01	< 0.01	> -0.01	< 0.01
	[0.006]	[0.006]	[0.005]	[0.006]	[0.005]	[0.005]
Asian: Prof	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	[0.003]	[0.002]	[0.002]	[0.003]	[0.002]	[0.002]
Hispanic: App	< 0.01	< 0.01	0.01	0.01	> -0.01	> -0.01
	[0.011]	[0.010]	[0.008]	[0.014]	[0.011]	[0.009]
Hispanic: Prof	> -0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	[0.005]	[0.004]	[0.003]	[0.004]	[0.003]	[0.003]
FRPL: App	0.12**	0.11**	0.07*	0.09	0.07	0.04
	[0.046]	[0.037]	[0.031]	[0.055]	[0.041]	[0.034]
FRPL: Prof	0.01	0.01	< 0.01	0.02	0.01	< 0.01
	[0.019]	[0.015]	[0.013]	[0.015]	[0.012]	[0.010]
ESL: App	-0.01	-0.01	-0.01	-0.02	< 0.01	< 0.01
	[0.012]	[0.010]	[0.009]	[0.015]	[0.012]	[0.010]
ESL: Prof	> -0.01	-0.01	> -0.01	-0.01	-0.01	-0.01
	[0.005]	[0.004]	[0.004]	[0.005]	[0.004]	[0.003]
Immigrant: App	0.02	0.01	< 0.01	0.01	0.01	< 0.01
	[0.016]	[0.014]	[0.011]	[0.018]	[0.015]	[0.012]
Immigrant: Prof	> -0.01	< 0.01	> -0.01	> -0.01	< 0.01	< 0.01
	[0.004]	[0.003]	[0.003]	[0.003]	[0.002]	[0.002]
	-0.02	0.04	0.08	0.05	0.10	0.12

Prior-Year Observation Score^{+}: App	[0.069]	[0.057]	[0.051]	[0.094]	[0.074]	[0.062]
Prior-Year Observation Score^{+}: Prof	-0.02	0.01	0.03	-0.07	-0.03	< 0.01
	[0.041]	[0.034]	[0.030]	[0.041]	[0.033]	[0.028]
N(Tch-Yrs)	3920	6015	8207	4228	6478	8750

Note: Estimates represent the total predicted change in the outcome. Standard errors, clustered at teacher-level, in brackets. OLS estimator employed to estimate all coefficients. BA+ is a binary variable indicating whether a teacher reported having a degree higher than a BA/ BS. Black is an indicator signaling whether the teacher reported her ethnicity/ race as Black or White. + Prior-year observation scores are not included as covariates in equations 1 and 2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3

Switching Job Assignments

	Math Teachers			Reading Teachers		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
Grade/ Subject Switch: App	-0.06	-0.08	-0.09	-0.31*	-0.29**	-0.18*
	[0.104]	[0.087]	[0.076]	[0.123]	[0.098]	[0.084]
Grade/ Subject Switch: Prof	-0.01	-0.03	-0.01	0.01	0.02	0.02
	[0.047]	[0.038]	[0.033]	[0.044]	[0.036]	[0.031]
Joint Significance: F-statistic	0.16	0.89	0.77	2.99	4.21*	2.45
N (Tch-Yrs)	3920	6015	8207	4228	6478	8750
Switch to Tested: App	0.08	0.14	0.14	-0.03	0.03	0.00
	[0.100]	[0.083]	[0.072]	[0.121]	[0.098]	[0.085]
Switch to Tested: Prof	0.02	-0.02	-0.04	0.01	-0.02	-0.03
	[0.043]	[0.036]	[0.031]	[0.042]	[0.034]	[0.030]
Joint Significance: F-statistic	0.38	1.56	2.44	0.05	0.20	0.65
N(Tch-Yrs)	3920	6015	8207	4228	6478	8750
Switch to New School: App	0.05	0.06	0.01	0.05	0.03	-0.03
	[0.083]	[0.069]	[0.059]	[0.084]	[0.066]	[0.056]
Switch to New School: Prof	0.03	0.03	0.02	> -0.01	0.02	0.01
	[0.033]	[0.027]	[0.023]	[0.028]	[0.023]	[0.020]
Joint Significance: F-statistic	0.54	1.11	0.49	0.18	0.54	0.37
N(Tch-Yrs)	3920	6015	8207	4228	6478	8750

Note: Estimates represent the total predicted change in the outcome. Standard errors, clustered at teacher-level, in brackets. OLS estimator employed to estimate all coefficients. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4

Impetus to Improve: Testing Joint Significance of Instruments

	$w = 20$	$w = 30$	$w = 40$
Sum: Svy Hrs in PD (*PDhrs*)	0.13	1.08	0.72
	(1698)	(2526)	(3318)
Sum: Svy Tch Collab (*tchcollab*)	1.57	0.85	1.12
	(709)	(1087)	(1439)
Sum: Svy Exerted More Effort (*effortsum*)	0.92	0.12	< 0.01
	(1084)	(1589)	(2046)
Sum: Svy Hrs Improved Instruction (*insthrs*)	0.13	0.18	0.64
	(2721)	(4181)	(5591)
Sum: Svy Hrs Prepped for Obs (*obshrs*)	0.89	0.76	0.21
	(6417)	(9417)	(12174)

Note: *p*-values in brackets, number of teacher-year records in parentheses. All models include teacher demographics, certification status, controls for the distribution of teacher effectiveness at the school level, second order polynomial of LOE interacted with teacher certification status, and year fixed effects. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5

Effects by Thresholds and Certification

	Math Teachers			Reading Teachers		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
2nd Stage: Number of Observations per Year	-0.12	-0.06	-0.01	-0.02	0.01	0.01
	[0.084]	[0.063]	[0.053]	[0.052]	[0.033]	[0.027]
App Below LOE 425	1.28**	1.10***	0.97**	0.88*	1.09**	0.92**
	[0.390]	[0.308]	[0.267]	[0.446]	[0.330]	[0.283]
Prof Below LOE 425	0.37***	0.39***	0.44***	0.52***	0.57***	0.58***
	[0.094]	[0.076]	[0.067]	[0.093]	[0.075]	[0.065]
F-statistic	9.97	18.01	27.52	23.50	47.43	64.54
N (Tch-Yrs)	3920	6015	8207	4228	6478	8750
	Professional Teachers Only					
2nd Stage: Number of Observations	-0.27	-0.09	-0.01	< 0.01	0.02	0.02
	[0.147]	[0.089]	[0.067]	[0.052]	[0.036]	[0.030]
1st Stage F-statistic	12.90	27.52	46.29	39.22	81.86	114.73
N (Tch-Yrs)	3291	5133	7083	3670	5678	7723

Note: Teacher-clustered standard errors in brackets. All models include a polynomial of the average student prior achievement scores for students taught in year *t* (for example, the 2011-12 scores of students taught in 2012-13), proportion of students taught exhibiting various characteristics, teacher demographics including certification status, controls for the distribution of teacher effectiveness at the school level, a second order polynomial of LOE-cont interacted with teacher certification status, and year fixed effects. First stage estimate represents the total effect of crossing the threshold. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 6

Longer-Term Effects of Observations

	Math Teachers			Reading Teachers		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
2nd Stage: Number of Prior-Year Observations	0.05	-0.03	0.02	-0.01	0.01	0.02
	[0.107]	[0.084]	[0.070]	[0.050]	[0.039]	[0.034]
1st Stage F-statistic	10.03	15.23	22.36	15.43	26.36	38.33
N(Tch-Yrs)	2589	3871	5200	2770	4106	5502
	Professional Teachers Only					
2nd Stage: Number of Prior-Year Observations	< 0.01	-0.08	-0.02	-0.04	-0.02	-0.03
	[0.113]	[0.089]	[0.072]	[0.074]	[0.054]	[0.044]
1st Stage F-statistic	18.94	28.73	42.32	22.32	39.18	59.68
N(Tch-Yrs)	2189	3300	4466	2395	3597	4845

Note: Teacher-clustered standard errors in brackets. Equations use twice-lagged instruments and running variables, but outcomes and controls are unchanged.

Table 8

Demoralization: Effects of Crossing LOE-cont 425 on Leaving Teaching

	$w = 20$	$w = 30$	$w = 40$
Apprentice: Crossing Prior-Year LOE-Cont 425	0.02	0.01	0.01
	[0.017]	[0.014]	[0.012]
Professional: Crossing Prior-Year LOE-Cont 425	0.01	0.01	< 0.01
	[0.006]	[0.005]	[0.004]
N (Tch-Yrs)	32891	49698	64026

Note: Teacher-clustered standard errors in brackets. The predictors of interest are crossing the LOE-cont 425-threshold for Apprentice and Professional teachers. Controls are unchanged.

Table 9

Effects of Crossing LOE at Other Thresholds

	Math Teachers			Reading Teachers		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
Crossing Prior LOE-Cont at 275	> -0.01	> -0.01	-0.01	0.01	0.01	< 0.01
	[0.038]	[0.030]	[0.026]	[0.023]	[0.019]	[0.017]
N (Tch-Yrs)	1822	2718	3548	2444	3701	4951
Crossing Prior LOE-Cont at 350	0.02	0.04	0.02	0.01	0.01	0.01
	[0.023]	[0.020]	[0.018]	[0.014]	[0.012]	[0.010]
N (Tch-Yrs)	2985	4447	5921	5046	7491	9822

Note: Teacher-clustered standard errors in brackets. The predictor of interest is crossing the LOE-cont 275 or 350 thresholds. Controls are unchanged. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 10

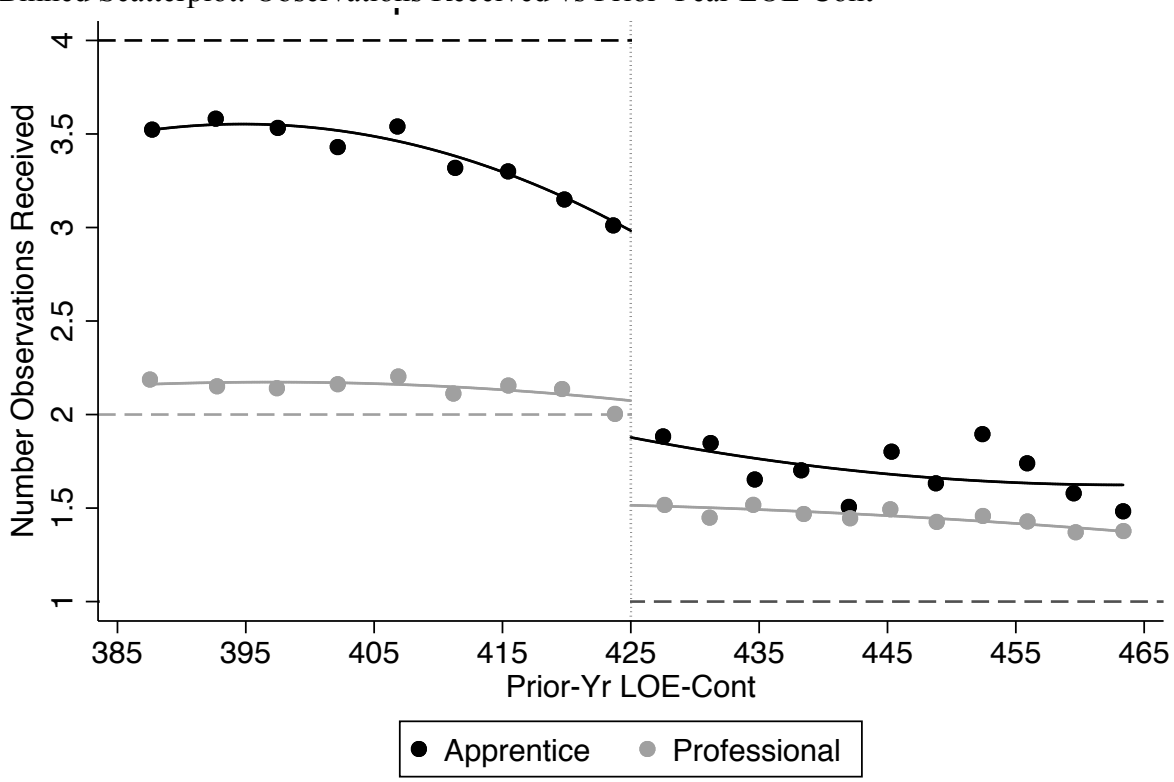
Implementation of TEAM Observation System

How much total time did you spend on pre-conferences? [2013, 2014]	0 hrs	< 1 hr	1-2 hrs	2-3 hrs	3-5 hrs	> 5 hrs
	18.4%	59%	16.8%	3.2 %	1.6%	1.1%
	(4,767)	(15,309)	(4,353)	(840)	(402)	(294)
How much total time did you spend receiving/ reviewing feedback from observations? [2013, 2014]	0 hrs	< 1 hr	1-2 hrs	2-3 hrs	3-5 hrs	> 5 hrs
	1.5%	68.3%	23.8%	4%	1.4%	1.1%
	(326)	(15,105)	(5,255)	(892)	(305)	(236)
My evaluator uses the rubric from our teacher evaluation process as a basis for discussing feedback from teaching observations. [2013, 2014]	Strongly Disagree	Disagree	Agree	Strongly Agree		
	2.9%	7.4%	58%	31.7%		
	(594)	(1526)	(11,885)	(6,496)		
I find it difficult to use feedback from my teaching observations to improve my practice. [2014]	Strongly Disagree	Disagree	Agree	Strongly Agree		
	6.7%	54.7%	32.2%	6.4%		
	(642)	(5,273)	(3,099)	(618)		
Did you take steps to address the indicator from your observations your evaluator identified as the one needing to be improved the most? [2013, 2014]	Yes	No				
	71.9%	28.1%				
	(7,635)	(2,981)				

Note: Survey responses from teachers with a prior-year LOE-cont $\geq$ 350. Survey items in first column. Years item administered on TES in brackets. Number of responses in parentheses.

Figure 1

Binned Scatterplot: Observations Received vs Prior-Year LOE-Cont



Note: Plotted points are the mean number of observations received within bins of four. Smoothed curves are second-order polynomials of the running variable, LOE-cont. A discontinuity in the number of policy-assigned observations exists at LOE-cont = 425. Horizontal dashed lines represent the policy-assigned number of observations. Policy assigns all teachers above 425 one observation, and Apprentice (Professional) teachers below 425 four (two) observations.

Figure 2

Distribution of Prior-Year LOE-Cont

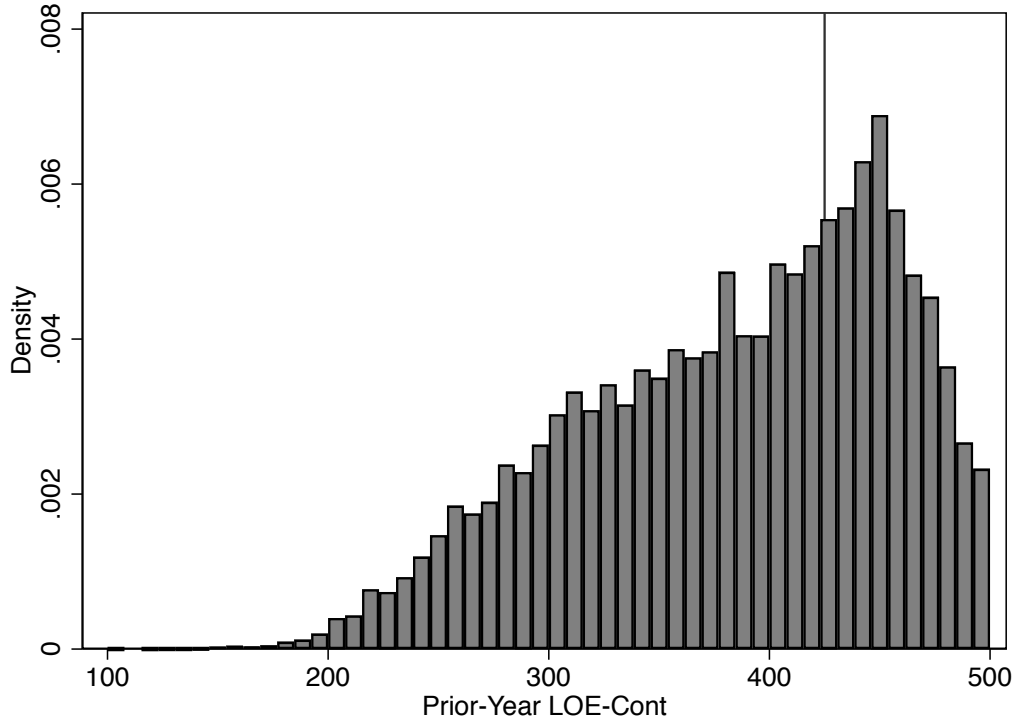
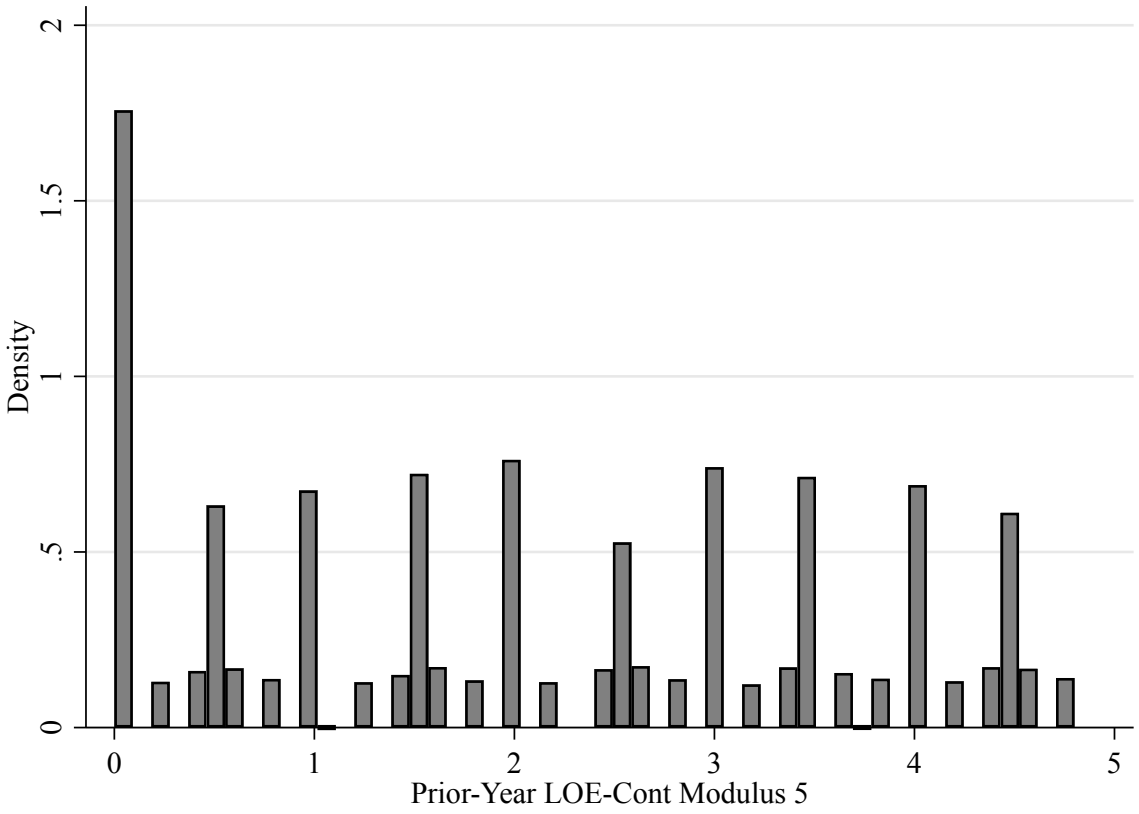


Figure 3

Distribution of Prior-Year LOE-Cont Transformed by Modulus Five



[i] State policy also assigns teachers with an LOE-cont $\leq$ 200 four observations, however, teachers in this range represent less than 0.75 percent of Tennessee teachers.

[ii] The Imbens-Kalyanaraman (cross-validation) bandwidth selector estimates an optimal bandwidth of 20 (75). The CV bandwidth is unreasonably large because the difference from one discrete LOE to the next is 75 on the LOE-cont scale.

[iii] These are polychoric correlations because growth and achievement scores are on an integer-scale. For more details about these measures see section 2.3 and Online Appendix 2.

[iv] Indeed, researchers using data from a different study context find evidence teacher performance affects the assignment of teachers to tested subjects (Grissom, Kalogrides, & Loeb, 2017).

[v] Others using Tennessee data also find no evidence that crossing LOE thresholds affect teacher professional development activities (Koedel, Li, Springer, & Tan, 2015).

[vi] When treating survey outcomes as ordinal or multinomial there was no evidence the proportional-odds assumption was valid and multinomial logit models failed to converge.

[vii] Means are found by taking the lower and upper bound of each response range. For example, I find the lower (upper) mean responses of "< 1 hr" by converting this response to one (59) minutes. The lower and upper values assigned to "0 hrs" are zero, and the upper value assigned to "> 5 hrs" is six.

References

Aaronson, D., Reserve, F., Barrow, L., Sander, W., Altonji, J., Butcher, K., … Diciccio, T. (2007). Teachers and Student Achievement in the Chicago Public High Schools, *25*(1).

Alexander, K. (2016). *TEAM Evaluator Training*.

Ballou, D., Canon, K., Ehlert, M., Wu, W. W., Doan, S., Taylor, L., & Springer, M. G. (2017). *Final Evaluation Report Tennessee's Strategic Compensation Programs Findings on Implementation and Impact 2010-2016*. Tennessee Consortium on Research, Evaluation, and Development.

Cattaneo, M., Jannson, M., & Ma, X. (2016). Simple Local Regression Distribution Estimators with an Application to Manipulation Testing.

Daley, G., & Kim, L. (2010). National Institute for Excellence in Teaching A Teacher Evaluation System That Works. *Working Paper*.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2017). Strategic Staffing? How Performance Pressures Affect the Distribution of Teachers Within Schools and Resulting Student Achievement. *American Educational Research Journal*, *54*(6), 1079–1116. https://doi.org/10.3102/0002831217716301

Guerin, B. (1993). *Social Facilitation* (1st ed.). Cambridge, UK: Cambridge University Press.

Halverson, R., Kelley, C., & Kimball, S. M. (2004). Implementing Teacher Evaluation Systems: How Principals Make Sense of Complex Artifacts to Shape Local Instructional Practice. In W. K. Hoy & C. G. Miskel (Eds.), *Educational Administration, Policy, and Reform: Research and Measurement*. Greenwich, CT: Information Age Publishing.

Jackson, C., & Cowan, J. (2018). *ASSESSING THE EVIDENCE ON TEACHER EVALUATION REFORMS* (Research Brief No. 13-1218–1) (p. 14). Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research.

Kimball, S. M. (2003). Analysis of Feedback, Enabling Conditions and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems. *Journal of Personnel Evaluation in Education*, *16*(4), 241–268. https://doi.org/10.1023/A:1021787806189

Koedel, C., Li, J., Springer, M. G., & Tan, L. (2015). *Do Evaluation Ratings Affect Teachers' Professional Development Activities?* (p. 57).

Kraft, M. A., & Gilmour, A. F. (2016). Can Principals Promote Teacher Development as Evaluators? A Case Study of Principals' Views and Experiences. *Educational Administration Quarterly*, *52*(5), 711–753. https://doi.org/10.1177/0013161X16653445

Manna, P. (2011). *Collision Course: Federal Education Policy Meets State and Local Realities*. CQ Press.

McGuinn, P. (2012). Stimulating Reform: Race to the Top, Competitive Grants and the Obama Education Agenda. *Educational Policy*, *26*(1), 136–159. https://doi.org/10.1177/0895904811425911

Mehta, J. (2013). *The Allure of Order*. Oxford University Press.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*. Thousand Oaks, CA: Sage Publications.

Neumerski, C. M., Grissom, J. A., Goldring, E., Cannata, M., Drake, T. A., Rubin, M., & Schuermann, P. (2014). Inside Teacher Evaluation Systems: Shifting the Role of

Principal as Instructional Leader. In *Association of Education Finance and Policy* (pp. 1–32). San Antonio, TX.

Olson, L. (2018). *SCALING REFORM: INSIDE TENNESSEE'S STATEWIDE TEACHER TRANSFORMATION* (p. 32). Washington, D.C.: FutureEd.

Papay, J. P., & Laski, M. E. (2018). *Exploring Teacher Improvement in Tennessee*. Nashville, TN: Tennessee Education Research Alliance.

Putman, H., Ross, E., & Walsh, K. (2018). *Making a difference: Six places where teacher evaluation systems are getting results*. Washington, D.C.: National Council on Teacher Quality.

Rigby, J. G. (2015). Principals' Sensemaking and Enactment of Teacher Evaluation. *Journal of Educational Administration*, *53*(3), 374–392. https://doi.org/10.1108/JEA-04-2014-0051

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement : Evidence from Panel Data. *The American Economic Review*, *94*(2).

Sartain, L., Stoelinga, S. R., Brown, E. R., Luppescu, S., Matsko, K. K., Miller, F. K., … Glazer, D. (2011). *Rethinking teacher evaluation: Lessons learned from observations, principal-teacher conferences, and district implementation*. Chicago, IL. Retrieved from https://consortium.uchicago.edu/sites/default/files/publications/Teacher Eval Report FINAL.pdf

Stecher, B. M., Garet, M. S., Hamilton, L. S., Steiner, E. D., Robyn, A., Poirier, J., … Brodziak de los Reyes, I. (2016). *Improving Teaching Effectiveness* (No. 9780833092212). Santa Monica, CA. Retrieved from

http://proxy.library.vcu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,cookie,uid&db=ehh&AN=11254922&site=ehost-live&scope=site

Steinberg, M. P., & Donaldson, M. L. (2016). The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era. *Education Finance and Policy*, *11*(3). https://doi.org/10.1162/EDFP_a_00186

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

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

Tennessee Board of Education. Teacher and Principal Evaluation Policy (2013).

US Department of Education. (2009). *Race to the Top Program Executive Summary*.

Walsh, K., Joseph, N., Lakis, K., & Lubell, S. (2017). *Running in Place: How New Teacher Evaluations Fail to Live Up to Promises*. National Council on Teacher Quality.

Online Appendix 1. Tennessee Educator Acceleration Model Observation Rubrics

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Standards and Objectives	• All learning objectives are clearly and explicitly communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are aligned and logically sequenced to the lesson's major objective. • Learning objectives are: (a) consistently connected to what students have previously learned, (b) known from life experiences, and (c) integrated with other disciplines. • Expectations for student performance are clear, demanding, and high. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.	• Most learning objectives are communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are mostly aligned to the lesson's major objective. • Learning objectives are connected to what students have previously learned. • Expectations for student performance are clear. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.	• Few learning objectives are communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are inconsistently aligned to the lesson's major objective. • Learning objectives are rarely connected to what students have previously learned. • Expectations for student performance are vague. • There is evidence that few students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.
Motivating Students	• The teacher consistently organizes the content so that it is personally meaningful and relevant to students. • The teacher consistently develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher regularly reinforces and rewards effort.	• The teacher sometimes organizes the content so that it is personally meaningful and relevant to students. • The teacher sometimes develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher sometimes reinforces and rewards effort.	• The teacher rarely organizes the content so that it is personally meaningful and relevant to students. • The teacher rarely develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher rarely reinforces and rewards effort.
Presenting Instructional Content	Presentation of content always includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • effective modeling of thinking process by the teacher and/or students guided by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • no irrelevant, confusing, or non-essential information.	Presentation of content most of the time includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • no irrelevant, confusing, or non-essential information.	Presentation of content rarely includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • relevant, coherent, or essential information.

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Lesson Structure and Pacing 	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end. • The lesson includes time for reflection. • Pacing is brisk and provides many opportunities for individual students who progress at different learning rates. • Routines for distributing materials are seamless. • No instructional time is lost during transitions.	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end. • Pacing is appropriate and sometimes provides opportunities for students who progress at different learning rates. • Routines for distributing materials are efficient. • Little instructional time is lost during transitions.	• The lesson does not start promptly. • The lesson has a structure, but it may be missing closure or introductory elements. • Pacing is appropriate for less than half of the students and rarely provides opportunities for students who progress at different learning rates. • Routines for distributing materials are inefficient. • Considerable time is lost during transitions.
Activities and Materials 	• Activities and materials include all of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking, o provide time for reflection, o are relevant to students' lives, o provide opportunities for student-to-student interaction, o induce student curiosity and suspense, o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • In addition, sometimes activities are game-like, involve simulations, require creating products, and demand self-direction and self-monitoring. • The preponderance of activities demand complex thinking and analysis. • Texts and tasks are appropriately complex.	• Activities and materials include most of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking; o provide time for reflection, o are relevant to students' lives, o provide opportunities for student-to-student interaction, o induce student curiosity and suspense; o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • Texts and tasks are appropriately complex.	• Activities and materials include few of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking, o provide time for reflection, o are relevant to students' lives, o provide opportunities for student to student interaction, o induce student curiosity and suspense, o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher made materials, manipulatives, resources from museums, etc.).

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)

Questioning 	• Teacher questions are varied and high quality, providing a balanced mix of question types:
	  ○ knowledge and comprehension,
	  ○ application and analysis, and
	  ○ creation and evaluation.
	• Questions require students to regularly cite evidence throughout lesson.
	• Questions are consistently purposeful and coherent.
	• A high frequency of questions is asked.
	• Questions are consistently sequenced with attention to the instructional goals.
	• Questions regularly require active responses (e.g., whole class signaling, choral responses, written and shared responses, or group and individual answers).
	• Wait time (3-5 seconds) is consistently provided.
	• The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex.
	• Students generate questions that lead to further inquiry and self-directed learning.
	• Questions regularly assess and advance student understanding.
	• When text is involved, majority of questions are text-based.
		• Teacher questions are varied and high quality providing for some, but not all, question types:
		  ○ knowledge and comprehension,
		  ○ application and analysis, and
		  ○ creation and evaluation.
		• Questions usually require students to cite evidence.
		• Questions are usually purposeful and coherent.
		• A moderate frequency of questions asked.
		• Questions are sometimes sequenced with attention to the instructional goals.
		• Questions sometimes require active responses (e.g., whole class signaling, choral responses, or group and individual answers).
		• Wait time is sometimes provided.
		• The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex.
		• When text is involved, majority of questions are text-based.
			• Teacher questions are inconsistent in quality and include few question types:
			  ○ knowledge and comprehension,
			  ○ application and analysis, and
			  ○ creation and evaluation.
			• Questions are random and lack coherence.
			• A low frequency of questions is asked.
			• Questions are rarely sequenced with attention to the instructional goals.
			• Questions rarely require active responses (e.g., whole class signaling, choral responses, or group and individual answers).
			• Wait time is inconsistently provided.
			• The teacher mostly calls on volunteers and high-ability students.

Academic Feedback 	• Oral and written feedback is consistently academically focused, frequent, high quality and references expectations. • Feedback is frequently given during guided practice and homework review. • The teacher circulates to prompt student thinking, assess each student's progress, and provide individual feedback. • Feedback from students is regularly used to monitor and adjust instruction. • Teacher engages students in giving specific and high-quality feedback to one another.	• Oral and written feedback is mostly academically focused, frequent, and mostly high quality. • Feedback is sometimes given during guided practice and homework review. • The teacher circulates during instructional activities to support engagement, and monitor student work. • Feedback from students is sometimes used to monitor and adjust instruction.	• The quality and timeliness of feedback is inconsistent. • Feedback is rarely given during guided practice and homework review. • The teacher circulates during instructional activities but monitors mostly behavior. • Feedback from students is rarely used to monitor or adjust instruction.

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Grouping Students 	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) consistently maximize student understanding and learning efficiency. • All students in groups know their roles, responsibilities, and group work expectations. • All students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) to best accomplish the goals of the lesson. • Instructional groups facilitate opportunities for students to set goals, reflect on, and evaluate their learning.	• The instructional grouping arrangements (either whole class, small groups, pairs, individual; heterogeneous or homogenous ability) adequately enhance student understanding and learning efficiency. • Most students in groups know their roles, responsibilities, and group work expectations. • Most students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) most of the time to best accomplish the goals of the lesson.	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) inhibit student understanding and learning efficiency. • Few students in groups know their roles, responsibilities, and group work expectations. • Few students participating in groups are held accountable for group work and individual work. • Instructional group composition remains unchanged irrespective of the learning and instructional goals of a lesson.
Teacher Content Knowledge 	• Teacher displays extensive content knowledge of all the subjects she or he teaches. • Teacher regularly implements a variety of subject-specific instructional strategies to enhance student content knowledge. • The teacher regularly highlights key concepts and ideas and uses them as bases to connect other powerful ideas. • Limited content is taught in sufficient depth to allow for the development of understanding.	• Teacher displays accurate content knowledge of all the subjects he or she teaches. • Teacher sometimes implements subject-specific instructional strategies to enhance student content knowledge. • The teacher sometimes highlights key concepts and ideas and uses them as bases to connect other powerful ideas.	• Teacher displays under-developed content knowledge in several subject areas. • Teacher rarely implements subject-specific instructional strategies to enhance student content knowledge. • Teacher does not understand key concepts and ideas in the discipline and therefore presents content in a disconnected manner.
Teacher Knowledge of Students 	• Teacher practices display understanding of each student's anticipated learning difficulties. • Teacher practices regularly incorporate student interests and cultural heritage. • Teacher regularly provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught.	• Teacher practices display understanding of some student anticipated learning difficulties. • Teacher practices sometimes incorporate student interests and cultural heritage. • Teacher sometimes provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught.	• Teacher practices demonstrate minimal knowledge of students anticipated learning difficulties. • Teacher practices rarely incorporate student interests or cultural heritage. • Teacher practices demonstrate little differentiation of instructional methods or content.

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)

Thinking 	• The teacher thoroughly teaches two or more types of thinking:
	  ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information;
	  ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios;
	  ○ creative thinking, where students create, design, imagine, and suppose; and
	  ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems.
	• The teacher provides opportunities where students:
	  ○ generate a variety of ideas and alternatives,
	  ○ analyze problems from multiple perspectives and viewpoints, and
	  ○ monitor their thinking to insure that they understand what they are learning, are attending to critical information, and are aware of the learning strategies that they are using and why.
		• The teacher thoroughly teaches one or more types of thinking:
		  ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information;
		  ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios;
		  ○ creative thinking, where students create, design, imagine, and suppose; and
		  ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems.
		• The teacher provides opportunities where students:
		  ○ generate a variety of ideas and alternatives, and
		  ○ analyze problems from multiple perspectives and viewpoints.
			• The teacher implements no learning experiences that thoroughly teach any type of thinking.
			• The teacher provides no opportunities where students:
			  ○ generate a variety of ideas and alternatives, or
			  ○ analyze problems from multiple perspectives and viewpoints.

Problem-Solving 	The teacher implements activities that teach and reinforce three or more of the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solutions • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing	The teacher implements activities that teach two of the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing	The teacher implements no activities that teach the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing

General Educator Rubric: Planning

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)

Instructional Plans ☐	Instructional plans include:
	• measurable and explicit goals aligned to state content standards;
	• activities, materials, and assessments that:
	  ○ are aligned to state standards,
	  ○ are sequenced from basic to complex,
	  ○ build on prior student knowledge, are relevant to students' lives, and integrate other disciplines, and
	  ○ provide appropriate time for student work, student reflection, and lesson unit and closure;
	• evidence that plan is appropriate for the age, knowledge, and interests of all learners; and
	• evidence that the plan provides regular opportunities to accommodate individual student needs.
		Instructional plans include:
		• goals aligned to state content standards,
		• activities, materials, and assessments that:
		  ○ are aligned to state standards,
		  ○ are sequenced from basic to complex,
		  ○ build on prior student knowledge, and
		  ○ provide appropriate time for student work, and lesson and unit closure;
		• evidence that plan is appropriate for the age, knowledge, and interests of most learners; and
		• evidence that the plan provides some opportunities to accommodate individual student needs.
			Instructional plans include:
			• few goals aligned to state content standards,
			• activities, materials, and assessments that:
			  ○ are rarely aligned to state standards,
			  ○ are rarely logically sequenced,
			  ○ rarely build on prior student knowledge, and
			  ○ inconsistently provide time for student work, and lesson and unit closure; and
			• little evidence that the plan provides some opportunities to accommodate individual student needs.

Student Work ☐	Assignments require students to: • organize, interpret, analyze, synthesize, and evaluate information rather than reproduce it, • draw conclusions, make generalizations, and produce arguments that are supported through extended writing, and • connect what they are learning to experiences, observations, feelings, or situations significant in their daily lives both inside and outside of school.	Assignments require students to: • interpret information rather than reproduce it, • draw conclusions and support them through writing, and • connect what they are learning to prior learning and some life experiences.	Assignments require students to: • mostly reproduce information, • rarely draw conclusions and support them through writing, and • rarely connect what they are learning to prior learning or life experiences.
Assessment ☐	Assessment plans: • are aligned with state content standards; • have clear measurement criteria; • measure student performance in more than three ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require extended written tasks; • are portfolio based with clear illustrations of student progress toward state content standards; and • include descriptions of how assessment results will be used to inform future instruction.	Assessment plans: • are aligned with state content standards; • have measurement criteria; • measure student performance in more than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require written tasks; and • include performance checks throughout the school year.	Assessment plans: • are rarely aligned with state content standards; • have ambiguous measurement criteria; • measure student performance in less than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); and • include performance checks, although the purpose of these checks is not clear.

General Educator Rubric: Environment

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Expectations 	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where all students can experience success. • Students take initiative and follow through with their own work. • Teacher optimizes instructional time, teaches more material, and demands better performance from every student.	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where most students can experience success. • Students complete their work according to teacher expectations.	• Teacher expectations are not sufficiently high for every student. • Teacher creates an environment where mistakes and failure are not viewed as learning experiences. • Students demonstrate little or no pride in the quality of their work.
Managing Student Behavior 	• Students are consistently well behaved and on task. • Teacher and students establish clear rules for learning and behavior. • The teacher overlooks inconsequential behavior. • The teacher deals with students who have caused disruptions rather than the entire class. • The teacher attends to disruptions quickly and firmly.	• Students are mostly well behaved and on task, some minor learning disruptions may occur. • Teacher establishes rules for learning and behavior. • The teacher uses some techniques, such as social approval, contingent activities, and consequences, to maintain appropriate student behavior. • The teacher overlooks some inconsequential behavior, but at other times, stops the lesson to address it. • The teacher deals with students who have caused disruptions, yet sometimes he or she addresses the entire class.	• Students are not well behaved and are often off task. • Teacher establishes few rules for learning and behavior. • The teacher uses few techniques to maintain appropriate student behavior. • The teacher cannot distinguish between inconsequential behavior and inappropriate behavior. • Disruptions frequently interrupt instruction.
Environment 	The classroom: • welcomes all members and guests, • is organized and understandable to all students, • supplies, equipment, and resources are all easily and readily accessible, • displays student work that frequently changes, and • is arranged to promote individual and group learning.	The classroom: • welcomes most members and guests, • is organized and understandable to most students, • supplies, equipment, and resources are accessible, • displays student work, and • is arranged to promote individual and group learning.	The classroom: • is somewhat cold and uninviting, • is not well organized and understandable to students, • supplies, equipment, and resources are difficult to access, • does not display student work, and • is not arranged to promote group learning.
Respectful Culture 	• Teacher-student interactions demonstrate caring and respect for one another. • Students exhibit caring and respect for one another. • Positive relationships and interdependence characterize the classroom.	• Teacher-student interactions are generally friendly, but may reflect occasional inconsistencies, favoritism, or disregard for students' cultures. • Students exhibit respect for the teacher and are generally polite to each other. • Teacher is sometimes receptive to the interests and opinions of students.	• Teacher-student interactions are sometimes authoritarian, negative, or inappropriate. • Students exhibit disrespect for the teacher. • Student interaction is characterized by conflict, sarcasm, or put-downs. • Teacher is not receptive to interests and opinions of students.

Online Appendix 2. Tennessee Measures of Teacher Effectiveness

Teacher Performance Measures Based on Student Outcomes

Two of three[viii] measures of teacher effectiveness are based on student outcomes: the achievement and growth scores. The achievement measure is a measure of district- / school-wide student outcomes including student achievement scores, graduation or attendance rates, etc. Teacher growth scores are based on student academic outcomes, but growth score options depend on whether the teacher teaches a tested subject (a "tested teacher").

A teacher and her school administrator[ix] choose an achievement measure at the beginning of each school year from a TDOE approved list of measures (Tennessee State Board of Education, 2013). Students in a teacher's school or district generate scores produced by each of these measures. Achievement measures are based on aggregations of grade-, department-, school-, or district-wide student outcomes (for example, achievement test scores). These aggregated measures are mapped onto an integer scale of one through five. (Tennessee Board of Education, 2013)

The second quantitative TDOE measure of teacher effectiveness is the "growth score". All teachers receive a growth score, however, the source of the score depends on whether the teacher teaches a tested subject or not. The Tennessee Value-Added Assessment System (TVAAS) estimates the impact of tested teachers on their students' test scores relative to the impact of the hypothetical average teacher on her students' test scores (SAS, 2016). TVAAS converts continuous value-added measures to an integer scale of one through five.

Teacher Performance Measures Based on Observational Ratings

Teacher observations in Tennessee are conducted by trained, certified[x] observers, 85% of whom are school administrators. Approximately half a teacher's observations are announced in advance. A complete observation cycle includes a pre-observation conference for announced observations, observation that may result in multiple scored domains (for example, Instruction, Environment), a post-observation feedback conference including design and/ or refinement of in/ formal teacher improvement plans.

Observers assign integer scores of one through five with respect to each indicator. The Instruction, Environment, and Planning domains have twelve, three, and four indicators, respectively. Exemplary teacher/ student behaviors with respect to each indicator are scored a five. The most undesirable behaviors receive a score of one. Behaviors "At Expectations" receive a three (see Online Appendix 1). If the observed evidence does not place a teacher squarely into one of the three levels of performance, an observer can assign a rating of two (four) for a preponderance of behaviors straddling the lowest and middle (and highest) categories. (Alexander, 2016)

[viii] Some Tennessee districts use a fourth LOE determinant: student perception surveys. These districts are excluded from the analysis because they use alternative observation systems.

[ix] Not all school administrators serve as teacher evaluators, nor are all teacher evaluators school administrators. Nevertheless, more than 85% of teacher evaluators are principals or assistant principals.

[x] TDOE hosts annual certification training. Participants must meet established performance expectations to become certified.

Online Appendix 3. Relationship Between Treatment and Observation Scores

Observation scores are susceptible to a source of bias that cannot affect student achievement: observer bias. Observers may rate teachers in such a way that confirms an un/ conscious impression: teachers in lower discrete LOE (i.e. LOE-cont below 425) are worse teachers and their observation scores should reflect this.

I hypothesize that if observer bias is present, observers will bring this bias into the classroom during their first observation of a teacher. Observer bias would cause the first observation score (hereafter "first score") a teacher receives to be lower, after controlling for the month of the observation, domains rated (that is, Instruction, Environment, Planning), and controls used by equations 1 and 2. It is plausible that the month during which a teacher receives their first observation is correlated with their performance (for example, observers may want to postpone difficult observations). It may also be the case that observers tend to rate teachers in one domain more harshly than another. Any effects on first scores cannot be due to genuine treatment effects because the teacher would not have received any post-observation feedback, or had time to respond to the first observation. Instead, estimates produced by my test of observer bias pick up the effect of a teacher's assignment to receive an additional observation. There is clear evidence suggesting observer bias exists. Table 3.1 shows the first score generated for teachers receiving an additional observation is systematically lower than teachers receiving fewer.

A thorough investigation of the sources of observer bias are beyond the scope of this paper, but such work would almost certainly be of interest to practitioners. By identifying the sources of bias, practitioners may be able to develop policies/ interventions that can mitigate the problem.

Table 3.1

Observer Bias

	DV = First Observation Score Received		
	w=20	w=30	w=40
2nd Stage: Number of Observations	-0.09	-0.10*	-0.08*
	(0.05)	(0.04)	(0.03)
App Below LOE-cont 425	1.13***	1.10***	1.20***
	(0.15)	(0.12)	(0.10)
Prof Below LOE-cont 425	0.57***	0.57***	0.59***
	(0.03)	(0.03)	(0.02)
N (Tch-Yrs)	22607	33546	43893

Notes: Standard errors clustered at teacher level. Each model controls for teacher demographics, school-level teacher effectiveness, LOE-cont, month of first observation, and domains scored on first observation. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Online Appendix 4. Tennessee Educator Survey Items

Table 4.1

Operationalized Variables	Survey Items	Scales	Years Administered	Operationalizations
Sum: Svy PD Hours (pdhrs)	Content: In-depth study of topics in my subjects	None; 1-5 Hours; 6-20 Hours; 21-40 Hours; More than 40 Hours	2013, 2014, 2015	Converted to hours by assigning each response to the lower bound of each item response interval (for example, the response "1-5" assigned to 1, response "More than 40" assigned to 40). Add together all responses across these items within a year to produce a PD hours sum.
	Preparing students to take the TCAP			
	Preparing students to take the CRA and/or writing assessments (2014, 2015 only)			
	Analyzing and interpreting student assessment results			
	Classroom organization			
	Teaching special student populations (e.g., English language learners and students)			
	Student behavior management			
	Addressing students' socio-emotional development and/or student behavior			
	Reviewing standards and curriculum to determine learning outcomes for my students			
	Pedagogy: Strategies for teaching my subject(s)			

Sum: Svy Tch Collab (tchcollab)	Met with other teachers to discuss standards, instruction, and/or student learning Met with the whole faculty at my school Worked with other teachers to develop materials or activities for particular classes Observed another teacher's classroom Reviewed student assessment data with other teachers to make instructional decisions	Never; About once a semester; About once a month; Two or three times a month; About once a week; More than once a week	2015	Responses converted into collaborative meetings per year. Never = 0, semesterly = 2, monthly = 7, two or three times monthly = 14, weekly = 28, more than weekly = 56. Add all responses across items within a year to produce a sum.
Sum: Svy Exerted More Effort (effortsum)	Focusing on the content covered by TCAP Engaging in other self-selected professional development opportunities to improve my content knowledge and/or teaching skills Reflecting on and discussing teaching and learning with my inquiry team or other teachers, coaches, etc. Tutoring individuals or small groups of students outside of class time Engaging in informal self-directed learning (e.g., reading a mathematics	Less time and effort than last year; The same time and effort as last year; More time and effort than last year; Not applicable	2013, 2014	Assigned "Less time" and "The same time" to 0, "More time" to 1. Add together all responses across these items within a year to produce a sum.

	education journal, using the Internet to enrich knowledge and skills)			
	Re-teaching topics or skills based on students' performance on classroom tests			
	Assigning or reassigning students to groups within my class			
	Preparing lessons			
	Differentiating instruction to address individual student needs			
	Communicating with parents orally or in writing			
	Attending district- or school-sponsored workshops			
	Integrating material from multiple subjects into lessons I teach (e.g., incorporating mathematics content into science or social studies classes)			
	Completing tasks required for teaching observations and evaluation activities			
	Disciplining students			
Svy Hrs Improved Instruction (insthrs)	Approximately how much time have you invested so far during the 2013-2014 school year in efforts to improve your instructional practices?	1-10 hours; 11-20 hours; 21-40 hours; 41-60 hours; 61-	2014	Converted to hours by assigning each response to the lower bound of each response interval

		80 hours; 81-100 hours; More than 100 hours		(for example, the response "1-10" assigned to 1, response "More than 100" assigned to 100)
Sum: Svy Hrs Prepped for Obs (obshrs)	How much TOTAL TIME have you spent on the following activities related to observations of your teaching during this school year?	None; Less than 1 hour; 1 to 2 hours; 2 to 3 hours; 3 to 5 hours; Over 5 hours	2013, 2014	Converted to hours by assigning each response to the lower bound of each response interval (for example, the response "Less than 1 hour" assigned to 0.5, response "1 to 2" assigned to 1, "Over 5 hours" assigned to 5)

Table 4.2

Descriptive Statistics. DV = Survey Items

	Mean	SD
Sum: Svy Hrs in PD	2.56	[17.46]
Sum: Svy Tch Collab	11.91	[41.10]
Sum: Svy Exerted More Effort	0.31	[1.54]
Svy Hrs Improved Instruction	37.27	[33.74]
Sum: Svy Hrs Prepped for Obs	2.25	[1.30]
Eval Will Improve Teaching	0.59	[0.492]
Post-Obs FB is Useful	0.87	[0.336]
Evals Change My Teaching	0.72	[0.449]
Observer Qualified	0.78	[0.413]
Evaluations Are Fair	0.64	[0.480]

Note: Descriptive statistics use data from the bandwidth 40 analytical sample. Standard deviations in brackets.

Online Appendix 5. Sensitivity of Instrument Validity to Treatment of Survey Measures

I use measures based on survey items to check for the presence of threats to the internal validity of the instruments and conclude the instruments are not threatened by an impetus to improve or demoralization. However, findings in Tables 4 may be sensitive to the treatment of survey items. I test the sensitivity of these findings using different operationalizations of the survey outcomes.

The original tests used five different measures based on survey items: hours spent in PD (*PDhrs*), frequency of teacher collaboration (*tchcollab*), hours spent improving instruction (*insthrs*), hours spent preparing for observations (*obshrs*), and effort invested in different activities (*effortsum*). The first four measures are based on survey items asking teachers about the amount of time spent on various activities. Each response to the original survey items[xi] is an interval/ frequency. For example, in response to items contributing to *PDhrs* teachers could choose: None, 1-5 Hours, 6-20 Hours, etc. The original operationalization of these items assigned each response the lower boundary of the chosen interval/ frequency. For example, if a teacher chose the "1-5 Hours" response this was converted to a value of one. The converted responses of items associated with *PDhrs* and *tchcollab* were then collapsed into single measures by adding all responses together, respectively. The *insthrs* and *obshrs* were each measured by a single item. Responses to original *effortsum* items were not intervals/ frequencies. These responses asked if a teacher exerted less, the same, or more effort on a particular activity. I originally assigned the lowest and middle categories to zero, and highest category to one, before collapsing all *effortsum* responses into a sum.

A new version of the *PDhrs*, *tchcollab, insthrs,* and *obshrs* items assigns original responses to the higher boundary of the chosen interval/ frequency. I refer to this as the *MAX*

conversion. A new version of *effortsum* items assigns the middle category to one instead of zero, which I refer to as the *HI* conversion. I also collapse multiple items measuring the same survey outcome (for example, time spent in professional development, time spent collaborating with colleagues) by taking means and sums. These new transformations yield eleven new outcomes: *PDhrsMAX, PDhrsmn, PDhrsMAXmn, tchcollabMAX, tchcollabmn, tchcollabMAXmn, effortsumHI, effortmn, effortHImn, insthrsMAX,* and *obshrsMAX.* All *MAX* (*HI*) measures are based on the *MAX* (*HI*) conversion and *mn* items are based on means instead of sums. Collapsed measures without "*mn"* in the label are based on sums.

Findings from these sensitivity tests are in Table 5.1. Results are qualitatively similar to those in Table 4: the instruments are unrelated to impetus to improve measures.

Table 5.1

Alternative Operationalizations of Impetus to Improve Outcomes

	$w = 20$	$w = 30$	$w = 40$
Sum: MAX Hrs in PD (*PDhrsMAX*)	0.37	1.11	0.56
	[0.693]	[0.329]	[0.572]
Mean: Hrs in PD (*PDhrsmn*)	0.15	0.95	0.45
	[0.860]	[0.385]	[0.638]
Mean: MAX Hrs in PD (*PDhrsMAXmn*)	0.46	0.95	0.31
	[0.629]	[0.386]	[0.732]
Sum: MAX Svy Tch Collab (*tchcollabMAX*)	1.58	0.89	0.94
	[0.208]	[0.412]	[0.392]
Mean: Svy Tch Collab (*tchcollabmn*)	1.61	0.81	1.27
	[0.201]	[0.443]	[0.280]
Mean: MAX Svy Tch Collab (*tchcollabMAXmn*)	1.61	0.85	1.04
	[0.201]	[0.429]	[0.352]
Sum: HI Svy Exerted More Effort (*effortsumHI*)	0.92	0.12	< 0.01
	[0.399]	[0.886]	[0.999]
Mean: Svy Exerted More Effort (*effortmn*)	0.81	0.05	< 0.01
	[0.444]	[0.950]	[0.996]
Mean: HI Svy Exerted More Effort (*effortHImn*)	0.81	0.05	< 0.01
	[0.444]	[0.950]	[0.996]
Sum: MAX Svy Hrs Improved Instruction (*insthrsMAX*)	0.06	0.15	0.40
	[0.937]	[0.857]	[0.670]
Sum: MAX Svy Hrs Prepped for Obs (*obshrsMAX*)	0.87	0.69	0.15
	[0.420]	[0.500]	[0.863]

Note: p-values in brackets. All models include teacher demographics, certification status, controls for the distribution of teacher effectiveness at the school level, a second order polynomial of LOE-cont interacted with teacher certification status, and year fixed effects. Samples sizes are the same as corresponding samples in Table 4. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

[xi] Despite the ordinal scale of these outcomes there is no evidence supporting the parallel regressions assumption.

Online Appendix 6. Timing of Observations

One explanation for the original null findings is that observers conduct the observations of teachers assigned more policy-imposed observations in bursts. To explore the hypothesis that the timing of observations accounts for the null results, I use "observation dates" in TDOE administrative data to find the fraction of all observations a teacher received over one- or two-month windows.

Observation dates in TDOE data capture the date an observer entered data into the Tennessee data management system, which is not necessarily the date of the associated observation. I assume the observation occurred in the month or prior-month of the observation date, however, TDOE is confident observers enter observation data within two to three weeks of an observation.

I check the sensitivity of results to different operationalizations. "Two-month Window A" pairs: August-September, October-November, December-January, February-March, and April-May. The "Two-Month Window B" construction pairs: July-August, September-October, November-December, January-February, March-April, May-June. Finally, I test the sensitivity of findings using one-month windows. All operationalizations generate similar results, and none of the results are significantly more positive than the main findings (see Table 8.1).

Table 6.1

Original Estimates and New Estimates After Accounting for the Timing of Observations

	Math			Reading		
	w = 20	w = 30	w = 20	w = 30	w = 20	w = 30
Original Estimates	-0.13	-0.09	-0.08	-0.07	-0.04	-0.06
Two-Month A: 95% CIs	[-0.33, 0.10]	[-0.29, 0.14]	[-0.30, 0.08]	[-0.41, 0.27]	[-0.22, 0.14]	[-0.22, 0.07]

Two-Month B: 95% Cis	[-0.39, 0.10]	[-0.40, 0.12]	[-0.30, 0.08]	[-0.90, 0.51]	[-0.39, 0.16]	[-0.39, 0.06]
One-Month: 95% CIs	[-0.38, 0.11]	[-0.39, 0.12]	[-0.45, 0.05]	[-0.86, 0.50]	[-0.38, 0.16]	[-0.39, 0.07]

Notes: Original and new estimates only differ in that models producing the new estimates account for the timing of observations received. Ninety-five percent confidence intervals in brackets. Standard errors clustered at teacher-level.