TN EDUCATION RESEARCH ALLIANCE

The Unintended Effects of Policy-Assigned Teacher Observations

Examining the Validity of Observation Scores

Seth B. Hunter

Abstract

Several state policies link high-stakes consequences to teacher evaluation scores, which tend to be heavily weighted by observation scores. However, research has only recently investigated the validity of these scores in modern teacher evaluation systems. I contribute to this body of work by examining the sensitivity of observation scores to a novel source of bias: the differentiated assignment of observation by state policy. Several states differentiate the number of observations assigned to teachers. Although the receipt of observations should influence observation scores, the differentiated assignment of observations to teachers should not. I apply a two-stage least squares regression discontinuity design to teacher panel data, exploiting discontinuities in Tennessee's differentiated assignment of observations, and find strong evidence of substantially negative bias. Multiple sensitivity tests strongly suggest that the assignment of observations by state policy is the source of assimilation bias, but this suggestion is not definitive. Implications are discussed.

WORKING PAPER
2019-05

The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores

A growing body of work examines the validity of teacher observation scores in modern evaluation contexts (e.g., Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016), and for good reason. Observation scores tend to receive the most weight among the multiple measures of teacher performance in teacher evaluation systems and, for many teachers, high-stakes consequences depend on evaluation scores (Cohen & Goldhaber, 2016; Steinberg & Donaldson, 2016). Nearly half of all states attach punitive consequences (e.g., loss of tenure, dismissal) to low teacher evaluation scores, and more than five link promotion and salary increases to evaluation results (American Institutes for Research, 2016). The amount of weight states typically give observation scores means these scores play a large role in the allocation of high-stakes consequences. Thus, substantial bias in observation scores (i.e., variation that is not attributable to teacher performance) would misallocate high-stakes consequences for reasons beyond a teacher's control.

Previous research concerning the predictors of observation scores primarily examines student performance and the characteristics of teachers and students (e.g., Campbell & Ronfeldt, 2018; Jacob & Walsh, 2011; Steinberg & Garrett, 2016). This work typically finds that observation scores are influenced or predicted by several student and teacher variables, such as student and teacher gender and race (Campbell & Ronfeldt, 2018), teaching experience (Jacob & Walsh, 2011), and the prior-year achievement of a teacher's incoming students (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). As the authors of these previous studies point out, these effects or relationships may or may not represent bias[1]. Instead, observation scores may capture intentional instructional adjustments or returns to teaching experience.

I extend this body of work by examining the sensitivity of observation scores to a novel source of bias: the number of observations assigned to teachers by state policy. More than thirty state education agencies currently differentiate the number of observations assigned to teachers (American Institutes for Research, 2016). Historically, states differentiating the assignment of observations did so on the basis of teaching experience or tenure (Steinberg & Donaldson, 2016). But in recent years, many states began differentiating observations based, in part, on prior-year teacher evaluation scores. Indeed, among the states differentiating observation assignment the majority assign observations on the basis of a teacher's prior-year evaluation scores (American Institutes for Research, 2016). Tennessee is among these states and is the context for this study.

At the start of each school year, Tennessee educators learn how many formal observations teachers should receive according to state policy. The theory of action undergirding modern observation systems implies that the *receipt* of observations might improve teacher performance (e.g., Georgia Department of Education, 2012; Tennessee Department of Education, 2016), and later observation scores may reflect improvement. However, the *assignment* of observations by state policy at the start of a school year should not affect observation scores outside the receipt of observations. For example, knowledge that policy assigns a teacher one observation, and another teacher two, should not affect how observers score teachers independent of observed performance. Yet, observers may believe teachers assigned more observations by state policy are worse teachers and (un)wittingly issue lower observation scores. Psychologists label this behavior "assimilation bias," because observers generate scores assimilating towards their perceptions of teacher prior performance, independent of observed performance (Sumer & Knight, 1996).

I find strong evidence suggesting that the assignment of the marginal observation by state policy lowers observation scores. Using longitudinal teacher-level data from more than 80 percent of Tennessee school districts, I exploit discontinuities in the Tennessee Board of Education's assignment of observations to explore bias brought about by this assignment. Estimates generated by regression discontinuity designs (RDDs) are substantially negative (though sometimes imprecise) at all thresholds where there is a discontinuity in the assignment of observations by state policy. Importantly, the assignment of observations by policy is a function of prior-year teacher effectiveness, raising questions about the source of bias. Evidence suggests that prior-year effectiveness does not generate the main findings, but this suggestion is not definitive. Nevertheless, there is clear evidence that some form of observer bias affects observation scores. Additional sensitivity tests rule out other threats to internal validity.

In the rest of this paper I discuss related literature and the Tennessee evaluation context. I then describe the methods and data used to generate the estimates of interest. After discussing the main findings, I present results from sensitivity tests. Considering the negative RDD estimates and what they could mean for practice, I explore the extent to which estimates depend on teaching experience and school administrator effectiveness/ skills (more than 85 percent of observers are school administrators). There is no evidence that the degree of bias depends on any examined moderator. The paper ends with conclusions, limitations, and implications.

Related Literature

Bias arising from the assignment of observations by policy is a form of *observer bias*, or what the broader literature characterizes as *rater bias*. Scholars define observer bias as the extent to which ratings systematically[2] deviate from a ratee's "true performance" (Bernardin, Thomason, Ronald Buckley, & Kane, 2016; Park, Chen, & Holtzman, 2015; Wherry & Bartlett,

1982). Observer bias has received a great deal of attention in the fields of psychology, educational measurement, and management, but few studies examine the degree of observer bias in modern teacher evaluation contexts. Studies examining this form of bias in modern evaluation contexts tend to focus on the influence of what I characterize as "teacher working conditions" and the characteristics of observers and/ or teachers (e.g., Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). This emerging body of work is important for reasons described below, but earlier psychological works imply other sources of bias beyond work conditions and observer/ ratee (teacher) characteristics exist. Specifically, features of observation systems may introduce observer bias.

In broad terms, previous work assigns observer bias to one of two categories, which I characterize as *context-independent* or *context-dependent* bias[3]. Context-independent bias originates from the observer herself and includes bias due to leniency, severity, and central tendency (Park et al., 2015). Observers who systematically issue lower (higher) ratings relative to other observers exhibit leniency (severity) bias (Engelhard, 1994). Instead of systematically generating scores towards the lower or upper end of rating scales, central tendency refers to the propensity of observers to issue ratings near the middle of the scale (Engelhard, 1994; Saal, Downey, & Lahey, 1980). A large number of studies in psychology and educational measurement focus on context-independent bias (e.g., Borman, 1975; Thorndike, 1920).

Relative to the body of work examining context-independent observer bias, context-*dependent* bias has received less attention (Park et al., 2015). However, there is growing interest in context-dependent sources of bias in teacher observation scores. Research suggests that teacher observation scores are influenced by several "working conditions" including grade taught, observer accountability, and student characteristics. For example, Mihaly and McCaffrey

find that observation scores vary by grade-level (2014). Observable teacher characteristics, characteristics of taught students, rater leniency/ severity, nor unobserved school-level factors explain these grade-level differences. The authors are unable to identify the source of grade-level variation, concluding it may exist because the observation instrument is better at capturing effective teaching practices in some grades more than others, or it may be difficult for observers to identify effective practices across grade levels.

Earlier work also suggests observation scores are affected by observer accountability. There is suggestive evidence that teacher observers generate more accurate ratings when they are accountable for issuing accurate scores (Graham, Milanowski, & Miller, 2012). Studies outside educational settings corroborate this finding (e.g., Wherry & Bartlett, 1982).

More recent work examines the influence of the classroom composition of students, which I characterize as a working condition. Steinberg and Garrett (2016), and Campbell and Ronfeldt (2018), use Measures of Effective Teaching (MET) project data to identify the extent to which researcher-generated observation scores produced in low-stakes settings are influenced by incoming student: achievement scores, race/ ethnicity, and gender. Steinberg and Garrett (2016) find that incoming achievement scores positively influence observational ratings; the researchers conjecture that higher observation scores may exist because of genuine teacher responses to their students or observer bias. Campbell and Ronfeldt (2018) find that teacher observation scores are lower when a higher proportion of the students taught are black, Hispanic, male, and have lower prior-year achievement scores. Campbell and Ronfeldt (2018) argue that differences in teacher ratings do not exist because of genuine differences in teacher quality, implying that these classroom characteristics are a source of observer bias.

Previous work also suggests that ratee (i.e., teacher) gender, race, and observer-ratee race-congruence generates observer bias. Using MET data, research finds male teachers receive systematically lower observation scores than females (Campbell & Ronfeldt, 2018). Research within and beyond educational settings finds black ratees (i.e., teachers) tend to receive lower ratings than whites (Arvey & Murphy, 1998; Campbell & Ronfeldt, 2018). And, ratees sharing the same race as their observer (i.e., race-congruent) tend to receive higher observational ratings than ratees who do not share the same race as their observer (Arvey & Murphy, 1998).

Psychologists argue that impressions of ratee prior performance are another source of observer bias, however, little empirical research has investigated this hypothesis (Wang, Wong, & Kwong, 2010). Observers who generate ratings resembling their impressions of ratee prior performance may exhibit what psychologists label *assimilation bias* (Sumer & Knight, 1996).

The existence of assimilation bias in high-stakes evaluation contexts has received little attention. Reviewed studies of assimilation bias occurred in university laboratory settings in which psychologists recruited undergraduate students to rate the performance of hypothetical employees (Sumer & Knight, 1996; Wang et al., 2010). These studies find some weak evidence of assimilation bias, but likely suffer from weak external validity, especially when considering if the findings generalize to high-stakes teacher evaluation systems. However, arguments advanced by psychologists suggest assimilation bias may exist in high-stakes teacher observation systems.

In this study I examine a potential source of assimilation bias in Tennessee teacher evaluation contexts, the number of observations assigned to teachers by state policy. Tennessee educators learn of this assignment at the beginning of each school year. The number of observations assigned to teachers by state policy may influence observer impressions of teacher performance, independent of observed teacher performance, introducing assimilation bias.

Moderators

Psychologists hypothesize that observer expertise in evaluation is negatively related to observer bias (e.g., Decotiis & Petit, 1978). A small body of work outside educational settings corroborates this hypothesis. In a review of literature outside educational settings, Graham, Milanowski, and Miller (2012) report that some work finds positive associations between observer expertise in evaluation and: rating accuracy and the ability of observers to differentiate among teacher behaviors across performance domains. Tziner, Murphy, and Cleveland (2005) also review related work outside educational settings and conclude that rater bias is a function of rater self-efficacy concerning performance management and evaluation. Finally, researchers using data associated with retail managers in a Fortune 500 firm find evidence supporting the hypothesis that expertise in evaluation partially accounts for rater bias (Bernardin et al., 2016).

This prior work suggests observer skills/ effectiveness moderates the degree of assimilation bias. I explore heterogeneous effects using four potential moderators based on measures of administrator[4] effectiveness/ skill adopted by the Tennessee Board of Education. Like teachers, administrators receive an average observation score based on a standards-based observation protocol. Administrators also receive an administrator effectiveness score ("LOE-cont") determined by administrator observation scores and student outcomes (see appendix). I treat 1) the measure of administrator effectiveness and 2) average administrator observation scores as proxies for generic administrator effectiveness and skill, respectively. The administrator observation rubric also includes indicators focused on 3) teacher professional learning and 4) teacher evaluation, which may be more direct measures of skills related to observer expertise. If true, the two latter measures may more strongly moderate the relationship between policy-assigned observations and first scores.

I also examine a fifth moderator: teacher years of experience. An established body of work finds positive returns to teaching experience in terms of student achievement growth (Harris & Sass, 2011; Ladd & Sorensen, 2017; Papay & Kraft, 2013) and observation scores (Jacob & Walsh, 2011). The positive returns to teaching experience may counter the negative effects of assimilation bias, should these effects exist.

Teacher Evaluation and Observation in Tennessee

The Tennessee Board of Education adopted the Tennessee Educator Acceleration Model (TEAM) in the early 2010s, introducing substantial changes to teacher evaluation and observation (Tennessee Board of Education, 2013). Changes included the introduction of a composite measure of teacher effectiveness ("Level of Effectiveness"), based on teacher observation scores and student outcomes. The Tennessee Board of Education also adopted a new observation schedule. Teacher Level of Effectiveness and *certification status* became the determinants of policy-assigned observations. In the Tennessee context, certification status is primarily a function of teaching experience.

Level of Effectiveness

TDOE generates two expressions of teacher Level of Effectiveness, which I label "LOE-cont" and "discrete LOE." LOE-cont is a continuous, composite measure of teacher effectiveness determined by observation scores and student outcomes[5]. Fifty percent of LOE-cont for teachers of tested subjects is based on student outcomes (e.g., value-added scores), the remaining half is based on observation scores. These same weights apply to teachers of untested subjects in the first study year (2012-13), but thereafter observation scores account for 60 percent of LOE-cont for teachers of untested subjects. TDOE converts LOE-cont to discrete LOE before sharing LOE with educators: LOE-cont scores within [100, 200), [200, 275), [275,350), [350, 425), or [425,

500] are respectively assigned discrete LOE performance categories of LOE1, LOE2, LOE3, LOE4, or LOE5 (Tennessee Department of Education, n.d.-b). During the study period (2012-13 through 2014-15), all educators received discrete LOE in the fall, before conducting or receiving observations. Educators did not receive LOE-cont scores.

TEAM Observation System Policy

During the study period, more than 80 percent of Tennessee districts adopted the state-default TEAM observation system. Considering its widespread adoption and (subsequently discussed) clear policies, this analysis focuses on the TEAM system. The Tennessee Board of Education adopted several policies concerning TEAM (Tennessee Board of Education, 2013). First, only annually certified observers may conduct formal classroom observations (henceforth "observations"). Annual observer certification focuses on generating accurate (i.e., unbiased) and reliable observation scores, and facilitating pre- and post-observation conferences aiming to improve teacher performance (Alexander, 2016). Second, TEAM districts must use the state-adopted TEAM rubric (see online appendix), which is based on the TAP rubric (Alexander, 2016). Observers use the TEAM rubric to measure teacher performance with respect to Planning, Instruction, and management of the (Classroom) Environment.

Third, at the beginning of each school year, the TENNESSEE BOARD OF EDUCATION assigns TEAM teachers observations based on their prior-year discrete LOE and current year certification status. Policy assigns teachers with a prior-year discrete LOE5 (LOE-cont $\geq$ 425) one observation, and teachers with a prior-year discrete LOE1 (LOE-cont < 200) four observations. The number of observations assigned to teachers with a prior-year discrete LOE2 through LOE4 (200 $\leq$ LOE-cont < 425) is two or four, depending on their current year certification status[6]. Certification status indicates whether a teacher taught less than four years

("Apprentice" teacher) or longer ("Professional"). Thus, there are two discontinuities in the number of policy-assigned observations at the LOE-cont 425-threshold, and one at the 200-threshold. According to THE TENNESSEE BOARD OF EDUCATION policy, districts/ schools/ educators can exercise discretion and conduct more than the policy-assigned number of observations, though no teacher should receive less than the number of policy-assigned observations (Tennessee Board of Education, 2013). Importantly, aside from the assignment of observations, crossing the 200- and 425-threshold does not trigger any other state policies[7].

Fourth, THE TENNESSEE BOARD OF EDUCATION has expectations concerning the timing and duration of observations. All teachers should receive their first observation in the fall. Teachers assigned more than one observation should receive approximately half their observations in the fall and the rest in the spring. The typical observation is expected to last at least 15 minutes.

Classroom Observations and the Improvement of Teacher Performance

The TEAM theory of action asserts classroom observations will improve teacher performance as measured by the TEAM observation rubric (Alexander, 2016; Tennessee Department of Education, 2016). The TEAM rubric describes a range of standards-based teacher behaviors, from unsatisfactory to exemplary behavior (Daley & Kim, 2010). Soon after each observation the observer should generate a TEAM score using the rubric. After generating this score, the observer is expected to meet with the teacher in a post-observation conference and supply feedback based on the TEAM score and rubric. TDOE expects observers to discuss teacher strengths and weaknesses, and work with teachers to develop improvement plans as needed. Improvement plans can take many forms, and may include teacher self-study, participation in workshops, coaching, etc.

Data and Methods

Data

I use teacher panel data from more than 80 percent of Tennessee school districts between the 2012-13 and 2014-15 school years. These data include teacher gender, level of education, race, years of experience, observation scores, and prior-year measures of LOE-cont. TEAM observation data include detailed teacher x observation-occurrence level records including observation dates and scores. Unique identifiers allow me to link teachers to schools.

Table 1 displays descriptive statistics for the population of Tennessee teachers, and teachers within the largest local RDD bandwidths at the LOE-cont thresholds of 200 and 425, respectively. The typical teacher in the Tennessee population is a white female with more than a bachelor's degree and approximately 12 years of experience. This typical teacher also receives approximately two observations per year and receives an average observation score of approximately four on the five-point ratings scale. The typical teacher in the largest bandwidth surrounding the 425-threshold resembles the typical Tennessee teacher, although there are fewer nonwhite teachers in the analytic sample. These similarities are unsurprising because, as shown at the bottom of Table 1, approximately 70 percent of teachers in the population have a discrete LOE4 or LOE5, the same discrete LOE straddling the 425-threshold.

In the bandwidth surrounding the 200-threshold, the typical Professional teacher resembles the typical teacher in the population with respect to: level of education, gender, and race. I restrict analytic samples at the 200-threshold to Professional teachers because there are no discontinuities in policy-assigned observations for Apprentice teachers at this threshold. Professional teachers in the largest bandwidth at the 200-threshold receive an observation score of approximately three on the five-point rating scale and, compared to teachers in the Tennessee

population, receive about one more observation per year and have almost two more years of experience. The difference in years of experience is expected because Professional teachers, by definition, must have at least four years of experience. Only nine percent of all Tennessee teachers in the study period straddle the 200-threshold, and less than half a percentage point of Tennessee teachers are assigned to a discrete LOE1. Approximately 12 percent of the analytic sample in the largest bandwidth surrounding the 200-threshold is below the threshold.

Some sensitivity analyses use data from the Tennessee Educator Survey (TES). Each spring, all Tennessee teachers receive the TES. Response rates exceed 50 percent.

Methods

At the start of each school year Tennessee educators learn how many observations teachers should receive, which may introduce assimilation bias. I test this hypothesis using 2SLS RDDs, exploiting information in teacher micro-data and plausibly exogenous discontinuities in the assignment of classroom observations by state policy.

A naïve empirical strategy might treat the average observation score of teacher *i* in year *t* (S_{it}) as the outcome of interest, and the total number of observations *received* over year *t* (g_{it}) as the predictor of interest, regressing S_{it} on g_{it} using OLS. This presents two significant problems, the first is using S_{it} as the outcome. The theory of action undergirding the TEAM observation system asserts that observation processes should improve S_{it}. To the extent the theory of action holds, estimates of the relationship between S_{it} and g_{it} may capture genuine teacher improvements brought about by *received* observation processes (g_{it}), in addition to observer bias, should it exist. However, received classroom observations cannot genuinely influence teacher performance prior to receipt of a post-observation conference because Tennessee observers do not share feedback or develop teacher improvement plans until the post-observation

conference. This implies that no observations received in year *t* can genuinely affect the *i*th teacher's first observation score (S_{1it}). I improve the naïve model by replacing S_{it} with S_{1it}.

A second, significant methodological problem remains. The total number of observations received (g_{it}) may still relate to S_{1it} (henceforth, "first scores") because the total number of observations *received* over the course of a year is potentially endogenous due to observer discretion. Observers may choose to observe less motivated teachers more often because the observer wants to closely monitor these teachers. Observers may (un)wittingly believe these teachers deserve lower observation scores, regardless of the performance seen during the first observation (negative bias). Alternatively, observers may avoid observing teachers who are unreceptive to post-observation feedback. Observers may (un)consciously conflate teacher reception of observer-provided feedback with teacher performance (positive bias). Either source of bias is troubling, but the estimates of interest in this study concern bias arising from the number of observations *assigned* by state policy[8].

To the extent observers conduct the number of observations assigned to teachers by state policy, I could estimate the relationship of interest by regressing S_{1it} on g_{it} using OLS or a one-stage RDD. However, educators do not "comply" with the policy-induced assignment of observations. Figures 1 and 2 show that some teachers receive more than the assigned number of observations over a school year, while others receive less. The dashed and dotted horizontal lines in Figures 1 and 2 represent the number of observations assigned to teachers by state policy at the start of the school year. As seen in Figure 1, policy assigns Professional teachers surrounding the 200-threshold four or two observations. Graphed points are the average number of observations received by teachers in bins of four on the LOE-cont scale, and the graphed lines capture trends in these graphed points. Figure 1 shows that Professional teachers below the 200-

threshold tend to receive the number of observations assigned by policy, but those above the threshold tend to receive 0.5 more observations than assigned by policy. A similar pattern exists among Professional teachers surrounding the 425-threshold (see Figure 2). However, Figure 2 shows that Apprentice teachers below 425 tend to receive about 0.5 *fewer* observations than expected, but those above 425 tend to receive about one more observation than assigned.

Differences between the number of observations received and assigned capture "non-compliance" with treatment assignment, where treatment is the number of observations assigned by state policy, and may represent endogenous variation that OLS or one-stage RDDs cannot plausibly overcome. In broad terms, the total number of observations received (g_{it}) is a function of educator discretion and policy-assignment. Because I can observe the determinants of policy-assignment, I use certification status and crossing the LOE-cont 425-threshold (200-threshold) as instruments for the total number of observations received in a 2SLS RDD. The following equations are estimated separately at the 200- and 425-thresholds.

$$1:\ g_{ikt} = \boldsymbol{\theta}\rho_{ikt} + \ddot{\boldsymbol{A}}h(\cdot) + \ddot{\boldsymbol{B}}\boldsymbol{X}_{ikt} + \ddot{\boldsymbol{C}}\boldsymbol{S}_{kt} + \gamma_t + \varpi_{ijt}, \qquad |LOEcont_{ikt}| \leq w$$

$$2:\ S_{1ikt} = \delta\widehat{g_{ikt}} + \boldsymbol{A}h(\cdot) + \boldsymbol{B}\boldsymbol{X}_{ikt} + \boldsymbol{C}\boldsymbol{S}_{ikt} + \gamma_t + \omega_{ijt}, \qquad |LOEcont_{ikt}| \leq w$$

Where S_{1ikt} is the first score received by teacher *i* in school *k* in year *t*, g_{ikt} the total number of observations received, ρ_{ikt} is a vector of two indicators signaling whether an Apprentice or Professional teacher is above or below the prior-year LOE-cont 425-threshold (200-threshold), $\widehat{g_{ikt}}$ the predicted number of observations from the first stage equation, and *h* is a second order polynomial of LOE-cont interacted with teacher certification status (i.e., Professional/ Apprentice). The relationship between *h* and the outcomes is allowed to vary across each threshold. $\boldsymbol{X}_{ikt}$ is a vector of covariates including teacher race/ ethnicity, gender, years of teaching experience, certification status, and level of education. I include these covariates to

increase precision. $\boldsymbol{X}_{ikt}$ also includes the month of the first observation and domains rated on the first observation. It is plausible that the timing of observations is correlated with teacher performance (e.g., observers may want to postpone difficult observations). It may also be the case that observers tend to score teachers in one domain more harshly than another[9]. $\boldsymbol{S}_{kt}$ is a vector of school level measures controlling for the distribution of teacher effectiveness in school *k* in year *t*, including the mean, standard deviation, and skewness of LOE-cont. Again, I include $\boldsymbol{S}_{kt}$ to improve precision. γ_t is a year fixed effect to account for secular trends, and ϖ_{ijt} and ω_{ijt} are idiosyncratic error terms. *w* represents bandwidths of 20, 30, and 40 on either side of the 425-threshold (200-threshold). *w* brackets the Imbens-Kalyanaraman (2012) optimal bandwidth[10]. Standard errors are clustered at the teacher level.

Equations 1 and 2 only include records associated with Professional teachers when using data from the 200-threshold. Including Apprentice teachers surrounding the 200-threshold increases statistical power but weakens the strength of the instruments (recall that Tennessee policy does not introduce a discontinuity for Apprentice teachers at this threshold).

δ, the relationship of interest, is estimated by comparing teachers falling just below the 200- (425-) threshold to those just above, within bandwidths *w*. To the extent identification assumptions are met, teachers fall just to either side of these thresholds according to a "locally random" process. Differences between the first scores of teachers falling just below a threshold, and those falling just above, can be attributed to crossing the threshold because this is the only systematic difference between these two groups. Because state policy assigns teachers below the 200- (425-) threshold more observations than those above, differences in first scores may be attributable to the assignment of observations by state policy. Importantly, state policy assigns all Tennessee teachers at least one observation per year. Thus, δ is based on comparisons of

teachers assigned x and $x + 1$ observations, where $x > 0$. If identification assumptions are met, δ is the local average treatment "effect" of assigning (not receiving) an additional policy-imposed observation on first scores. δ is local because it is estimated in bandwidths w and based on three policy-imposed discontinuities.

Even if teachers are locally randomized around the 200- and 425-thresholds, there are still threats to interpreting δ as the marginal effect of assigning a teacher an additional observation. The most obvious threat stems from the fact that crossing from above to below either threshold assigns teachers to a lower discrete LOE, which may introduce a confounding "alternative treatment." That is, crossing either threshold assigns teachers more observations (one treatment) and assigns them to a lower category of performance (a potential alternative second treatment). Assignment to a lower discrete LOE (ρ_{it}) may also introduce assimilation bias, directly affecting first scores (S_{1it}). Observers may (un)consciously believe teachers with lower prior-year discrete LOE deserve lower observation scores regardless of observed performance. To the extent assimilation bias exist at the 425- or 200-thresholds, the existence of an alternative treatment would make it difficult to disentangle the source of bias.

There is strong evidence of assimilation bias arising from the assignment of observations by state policy at the 200- and 425-thresholds (although 200-threshold estimates are imprecisely estimated). Sensitivity tests suggest that assignment to a lower discrete LOE does not generate the findings, although I am unable to definitively rule out this suggestion. After presenting the main findings, I discuss additional potential threats to my conclusion that: strategic rating by observers, easier scoring of teachers receiving a single observation (i.e., discrete LOE5 teachers), changes in teaching assignment, or teacher improvement efforts account for what I interpret as

assimilation bias. There is no evidence any of these alternative explanations account for my interpretation of the main findings.

Threats to Internal Validity

Manipulation of the Running Variable

The only individuals who could manipulate LOE-cont scores are observers. One may be concerned observers place teachers to either side of the 200- or 425-thresholds for reasons related to teacher performance. For example, an observer may try to place a teacher above a threshold because they believe the teacher is on a path towards improvement and does not need more policy-assigned observations. Such nonrandom assignment violates the RDD identification assumption of local randomization at the 200- and 425-thresholds.

However, manipulation of LOE-cont is practically infeasible. Observers do not receive the student outcomes that determine LOE-cont until after the completion of all observations. Thus, observers would need to accurately predict the student-outcome determinants of LOE-cont to engage in manipulation. Observers could turn to historic student outcomes to predict contemporaneous scores. However, the correlation between prior-year and contemporaneous student-based outcomes[11] is below 0.50. Despite the *ex-ante* infeasibility of manipulation, I devise a statistical test for manipulation under the assumption of observer prescience. If there is no evidence of manipulation under this assumption, there is little reason to believe observers manipulated LOE-cont under real conditions.

LOE-cont is only approximately continuous, invalidating the use of conventional tests for manipulation. Conventional tests of manipulation compare the probability density function (PDF) of the running variable as it approaches a threshold from the left to the PDF of the running variable as it approaches the threshold from the right. A relatively large difference between PDF

estimates is evidence of manipulation. However, this type of discontinuity in LOE-cont is expected. LOE-cont is a linear combination of observation, "achievement," and "growth scores" (see endnote 3). Achievement and growth scores are on an integer scale, but observation scores include rational numbers. The least common multiple of weights applied to these three scores is five. Thus, deviations from LOE-cont multiples of five only occur due to observation scores and there are "naturally" occurring discontinuities in the probability density function of LOE-cont at multiples of five, invalidating conventional tests of manipulation. Graphical evidence shows expected patterns in LOE-cont. Figure 3 is a histogram of prior-year LOE-cont. Figure 4 is the distribution of these same scores transformed via modulus five. Figure 4 shows zero is the modal LOE-cont score modulus five, meaning most LOE-cont scores are multiples of five.

To circumvent this problem, I assume observers are prescient, remove achievement and growth scores from LOE-cont, then test for manipulation. Specifically, I subtract the weighted achievement and growth scores from 425. This difference is the weighted observation score needed to produce an LOE-cont score of 425. I then subtract this difference from the weighted observation scores observers generated, producing what I characterize as *prescient* LOE-cont. A *prescient* LOE-cont score of zero would mean that the observer generated the exact observation score needed to give the teacher an LOE-cont of 425. I create an analogous version of *prescient* LOE-cont "centered" at the LOE-cont 200-threshold.

I apply the robust-bias correction approach to test for manipulation at *prescient* LOE-cont values of zero (Cattaneo, Jannson, & Ma, 2016). The robust-bias corrected approach does not reject the null hypothesis of no manipulation at zero on the *prescient* LOE-cont scales centered at LOE-cont values of 200 or 425. *p*-values generated by epanechnikov and triangular kernel functions of *prescient* LOE-cont corresponding to the LOE-cont 200-threshold are, 0.60 and

0.35, respectively. *p*-values produced by epanechnikov and triangular kernel functions of *prescient* LOE-cont corresponding to the 425-threshold are 0.17 and 0.11, respectively. Because there is no evidence of manipulation under conditions of observer prescience, I conclude that manipulation under realistic conditions is implausible.

Covariate Balance Tests

Local randomization is an identification assumption of RDDs. Covariate balance tests explore this assumption by testing if there are discontinuities in observable baseline teacher characteristics. Discontinuities in baseline characteristics at a threshold would threaten the RDD.

There is no evidence of discontinuities in baseline characteristics at the 200- or 425-threshold. The left and right panels of Table 2 displays results from covariate balance tests at the 200- and 425-thresholds, respectively. Professional teachers just to either side of the 200-threshold are indistinguishable with respect to: teaching experience, gender, degree, and race. At the 425-threshold, similar patterns exist among Professional and Apprentice teachers.

These results, combined with the discussion of manipulation, support the identification assumption that teachers fall just above or below the 200- and 425-thresholds according to a locally random process. However, teachers just below each threshold differ from those just above in two important ways. The former are assigned to a lower discrete LOE and assigned more observations at the start of the school year. To the extent there are differences in first scores, discussions in this section imply that these differences can plausibly be attributed to the assignment of observations by state policy and/ or assignment to a lower discrete LOE. After discussing the main results, I explore the extent to which one or both sources accounts for the main results. The evidence suggests that the assignment of observations by state policy generates the results, but this suggestion is not definitive.

Results

At each threshold, the assignment of an additional observation by state policy is negatively related to first scores, although estimates at the 200-threshold are imprecisely estimated (see Table 3). Across all bandwidths at the 425-threshold, the instruments strongly predict the endogenous regressor and first-stage F-statistics far exceed the benchmark value of ten. Estimates in the right panel of Table 3 suggest that the assignment of one more policy-induced observation lowers first scores by approximately 0.14 units on the observation rubric scale, or about 0.22 SD, for teachers in bandwidths *w*. Because the outcome of interest is first scores, this relationship does not capture the genuine influence of *receiving* observations on observation scores.

The left panel of Table 3 displays first- and second-stage estimates using data from the 200-threshold. F-statistics in bandwidths of 20 and 30 around the 200-threshold are near ten, and the F-statistic in the largest bandwidth is almost 14. At the 200-threshold, the assignment (not receipt) of an additional observation by state policy is predicted to lower first scores by about 0.22 units, or 0.30 standard deviations, a substantially large change in observation scores. However, all second-stage estimates at the 200-threshold are imprecisely estimated.

Given the imprecision of estimates at the 200-threshold, remaining discussions focus on estimates from the 425-threshold. To be clear, the lack of statistical significance does not refute the existence of assimilation bias at the 200-threshold. Indeed, the large negative relationships suggest that, if anything, assimilation bias could be detected with more statistical power.

Sensitivity Tests

Estimates at the 425-threshold suggests that some form of assimilation bias is present. Although there is evidence of assimilation bias it is unclear if the estimates are driven by: the

number of observations assigned to teachers by state policy, or assignment to a lower discrete LOE. Either source of bias is problematic, but knowing the source would allow policymakers or education agencies to address the primary source. The evidence strongly suggests that the number of observations assigned by state policy drives the main findings, but this suggestion is not definitive.

There are four more alternative explanations that may account for the main results. First, features of the administrator evaluation system may unintentionally incentivize observers to strategically issue teachers in lower discrete LOE lower first scores, such that teachers in lower discrete LOE appear to "grow" within a school year [12]. Such strategic scoring may account for the main findings. The second alternative explanation is that observers may score teachers *receiving* only one observation more leniently than others because this group of teachers will not have an opportunity to show later within-year growth. This practice would inflate the observation scores of teachers just above the 425-threshold, which may explain the main findings. Third, principals may reassign higher performing (i.e., LOE5) teachers of untested subjects to tested subjects (Grissom, Kalogrides, & Loeb, 2017), introducing instructional challenges that lower observation scores. The final alternative explanation concerns teacher reactions to receiving an LOE4 instead of an LOE5, independent of observation processes. Teachers may interpret assignment to a lower discrete LOE as a signal of failure and abandon improvement efforts, accounting for the negative effects on first scores. No sensitivity tests substantiate any of these threats, suggesting that the main findings represent assimilation bias introduced by policy-assigned observations.

Assimilation Bias Arising from Prior-Year Discrete LOE. I explore the extent to which assignment to a lower discrete LOE generates the main findings by estimating the effect

of assignment to a lower discrete LOE at the LOE-cont 275- and 350-thresholds where there are no discontinuities in policy-assigned observations. These tests explore the presence of a "generic" relationship between prior-year discrete LOE and observation scores. To the extent a generic negative effect between prior-year discrete LOE and first scores exists, evidence to this effect should appear at the 275- and 350- thresholds. Because crossing either of these thresholds only assigns teachers to a lower discrete LOE, and not more observations, I can compare estimates driven by a generic discrete LOE effect to those that may be driven by a generic LOE effect *and* assignment to more observations (i.e., 425-estimates).

There is no evidence that assignment to a lower discrete LOE at the 275- or 350-thresholds negatively affects observation scores (see Table 4). Crossing the 275-threshold has a statistically insignificant, near-zero effect on first scores, and crossing the 350-threshold tends to have a positive effect. These near-zero and positive relationships suggest that assignment to LOE4 instead of LOE5 has a non-negative effect on observation scores, implying that the assignment of observations by state policy generates the main findings. Although the evidence in Table 4 is compelling, it is not definitive. At the 425-threshold, I cannot definitively disentangle the influence of assignment to LOE4 from the assignment of more observations by state policy.

Strategic Low-to-High Scoring. More than 85 percent of observers are school administrators who are expected to develop teacher performance. School administrators may respond to this expectation by strategically scoring teachers assigned more observations lower on the first observation, only to rate them higher on later observations. Doing so would suggest the school administrators' relatively less effective (i.e., LOE4) teachers experience within-year "growth." Such "growth" would be captured in teacher observation scores, which feeds into her LOE. Importantly, such "growth" can also affect measures of school administrator effectiveness.

The TEAM evaluation system includes teacher and school administrator evaluation systems (Tennessee Board of Education, 2013). Like teachers, school administrators receive an LOE based on student outcomes and "observation scores." However, administrator observation scores are not based on live-observation of administrator performance (Tennessee Department of Education, 2016). Instead, administrator evaluators can use administrator-generated teacher TEAM scores as a partial determinant of administrator observation scores[13]. Thus, it is possible for school administrators to (un)consciously influence their own observation scores.

I explore low-to-high scoring by replacing the first score in equation 2 with the difference between a teacher's first score and their mean observation score $(S_{1ikt} - \bar{S}_{\cdot ikt})$, otherwise, equations 1 and 2 remain unchanged. A large positive (negative) value of the new mean-centered outcome shows the first score is "atypically" higher (lower) than the teacher's own mean observation score. If the mean-centered first scores for teachers assigned more observations is significantly lower, this could suggest low-to-high scoring practices.

There is no evidence that teachers assigned more observations by state policy receive more atypical first scores compared to their own mean observation score. The main findings revealed that the first scores of teachers assigned an additional observation are approximately 0.14 units lower. However, results in Panel A of Table 5 show that mean-centered first scores do not vary with policy-assigned observations: estimates are near-zero and statistically insignificant. This implies that the assignment of another observation by state policy reduces first and mean observation scores by approximately 0.14. Therefore, results in the top panel of Table 5 suggest that strategic low-to-high scoring practices do not generate the main findings.

Sensitivity to Easier Scoring of Teachers Receiving One Observation. Observers may score teachers *receiving* a single observation higher than other teachers. Observers may judge

these teachers more leniently because this group of teachers does not have an opportunity to exhibit better performance later in the year. I explore this hypothesis by estimating effects using a 425-threshold sample restricted to teachers *receiving* more than one observation. This restriction reduces the original 425-threshold samples by approximately 30 percent, doubling to tripling the original standard errors. Nonetheless, Panel B in Table 5 shows there is a negative relationship between the assignment of an additional observation by state policy and first scores, and that the new point estimates resemble the size of the original 425-threshold estimates. Despite the imprecision of the new estimates, there is no evidence observers score teachers receiving only one observation easier than teachers receiving more observations.

Switching Teachers to (Un)Tested Subjects. Prior work finds that principals assign higher performing teachers to tested subjects (Grissom et al., 2017). Assignment to a new position may introduce new instructional challenges, especially during a teacher's first year in a new position. To the extent these challenges exist, they may lower observation scores, introducing negative bias. I examine changes in teaching a(n) (un)tested subject by creating a dummy variable taking a value of one if a teacher switched from a tested subject in year $t-1$ to an untested subject in t, or from an untested subject in year $t-1$ to a tested subject in t. The dummy takes a value of zero if the teacher remained in a(n) (un)tested subject in years $t-1$ and t. I regress this dummy on all right-hand side variables in equation 1.

There is no evidence that crossing the 425-threshold affects teaching assignments with respect to teaching a tested subject (see Table 5 Panel C). Crossing from the 425-threshold tends to have a near-zero effect across all bandwidths for Apprentice and Professional teachers.

Abandonment of Improvement Efforts. Assignment to LOE4 may induce teachers to abandon improvement efforts, independent of observational processes, negatively affecting

performance. Evidence to this effect should appear in teacher responses to TES (survey) items concerning their improvement efforts. The TES asked teachers about their: engagement in professional development[14], instructional improvement efforts, and exertion of effort on various activities (e.g., lesson preparation, reflecting on teaching). The online appendix describes these items, the original scales, descriptions of the transformations of these items into the outcomes of interest, and descriptive statistics. Evidence that teachers below the 425-threshold report lower (i.e., more negative) improvements efforts would threaten my interpretation of the main findings.

I examine the effect of crossing the 425-threshold on survey outcomes by regressing[15] survey outcomes on the right-hand side variables from equation 1. When Professional teachers are assigned to LOE4 in a bandwidth of 20, there is a statistically significant drop in the survey outcome regarding the exertion of teacher effort. However, this effect is only detected at the five percent level and, given the number of tests conducted, this estimate may exist due to Type I error. Moreover, the Apprentice and Professional indicators in this bandwidth are not jointly significant (see Table 6). No other effects are individually or jointly significant. I conclude that teacher abandonment of improvement efforts does not threaten my interpretation of the results.

Moderation Analyses

To explore heterogenous effects, I convert each of the four measures of administrator skill/ effectiveness to school x year quartiles[16]. The teaching experience moderator is a teacher x year measure. Some moderated relationships follow hypothesized patterns, but none of the differences across quartiles are substantively or statistically significant.

Panel A of Table 6 displays relationships moderated by quartiles of administrator LOE-cont. Across all bandwidths, the magnitude of negative estimates tends to decrease as administrator LOE-cont scores rise, with the clearest trend in a bandwidth of 40. However, none

of the relationships across any two quartiles are statistically different at conventional levels (i.e., the 95 percent confidence intervals of any two quartiles overlap).

Panel B of Table 6 displays estimates moderated by administrator average observation scores. As administrator average observation scores rise from the first through third quartiles, the negative estimates attenuate (though differences are statistically insignificant). However, the size of the relationship increases among the highest rated administrators. Results in Panel A of Table 7 show relationships moderated by administrator skills as teacher evaluators. These results resemble the pattern of relationships moderated by average observation scores.

Relationships moderated by administrator skills regarding teacher professional learning provide conflicting evidence (see Panel B of Table 7). The smallest degree of assimilation bias is found in the second quartile of administrator skills concerning teacher professional learning, contrary to expectations. However, the pattern of relationships across the first, third, and fourth quartiles support the moderating hypothesis. Moreover, the most and least negative relationships are found in the first and fourth quartiles, respectively.

Teaching experience does not moderate the degree of assimilation bias. Panel A of Table 8 shows the main effect of observation assignment, and interaction between observation assignment and a linear measure of teaching experience. The interaction is near-zero and statistically insignificant. I also create a three-category binned form of teaching experience (1-2 years, 3-5 years, 6 or more years) in the event teaching experience moderates assimilation bias according to a non-linear relationship. Panel B of Table 8 displays the non-linear moderation analysis, in which the reference group is 1-2 years of experience. Interactions between assigned observations and binned years of experience are also near-zero and statistically insignificant.

Despite the existence of some weak supportive evidence, I conclude that these five measures do not moderate the relationship between first scores and the number of observations assigned by state policy.

Discussion

At the start of each school year, Tennessee observers learn how many formal observations teachers have been assigned by state policy. Previous work concerning "assimilation bias," the (un)conscious tendency of observers to generate observations scores resembling their impressions of employee performance independent of observed employee performance (Sumer & Knight, 1996), implies that the number of observations assigned by state policy might bias observation scores. Observers may (un)wittingly issue lower observation scores to teachers assigned more observations because observers (un)consciously believe worse teachers are assigned more observations.

2SLS RDDs find strong evidence of assimilation bias. The first observation score of teachers in bandwidths surrounding RDD thresholds is predicted to decline by at least 0.20 SD with the assignment of an additional observation by state policy. Although predicted declines at the 200-threshold are imprecisely estimated, the declines exceed 0.30 SD in first observation scores. Additionally, moderation analyses suggest neither teaching experience nor the measures of administrator skill and effectiveness adopted by the Tennessee Board of Education moderate the degree of assimilation bias.

To place the degree of bias in context, previous work suggests that the average observation scores received by male teachers is 0.10 to 0.20 SD lower than female teachers, and average observation scores are predicted to decline by approximately 0.10 SD if the proportion of a teachers' students who are black increases by 25 percent (Campbell & Ronfeldt, 2018). In

terms of the characteristics of Tennessee teachers, 0.22 SD is approximately half the difference in average observation scores between first- and second-year teachers. Thus, the main findings capture substantial changes in observation scores.

Several ancillary analyses support RDD identification assumptions and strongly suggest that the source of assimilation bias at the 425-threshold is the assignment of observations by state policy. Tests for manipulation of the running variable and covariate balance support the identification assumption of "local randomization" at this threshold. Additional sensitivity tests find no evidence that plausible alternative explanations account for the estimated relationships.

The most plausible threat to interpreting policy-assigned observations as the source of assimilation bias is the potential existence of what I characterize as an "alternative treatment" at the 425-threshold. Crossing from just above to just below the threshold results in the assignment of more observations by state policy (i.e., treatment of interest) and assigns teachers to a lower category of teacher effectiveness (i.e., potential alternative treatment). Observers may (un)wittingly introduce assimilation bias because they believe teachers in a lower discrete LOE deserve lower observation scores, not because state policy assigns these teachers more observations. Yet, no evidence suggests that assignment to a lower discrete LOE generates findings at the 425-threshold. Notwithstanding this strong suggestion, I cannot definitively conclude that assignment to a lower discrete LOE is not a source of assimilation bias.

Limitations

There are three potential limitations to this study. First, estimates at the 200- and 425-threshold are based on variation from records just to either side of these thresholds which may limit the generalizability of these findings with respect to teacher effectiveness. Moreover, the only statistically significant main findings come from data surrounding the 425-threshold (i.e.,

teachers assigned to one of Tennessee's top two categories of teacher effectiveness). It is also the case that the majority of identifying variation at the 425-threshold is associated with the marginal change from one to two policy-assigned observations. However, more than 40 percent of all Tennessee teachers are in the largest bandwidth of 40 and the marginal change from one to two policy-assigned observations is the margin applying to most Tennessee teachers. Moreover, this is the margin applying to many teachers across the United States. Based on policies from 46 states, the typical (i.e., mean) beginning teacher and career teacher is assigned approximately two formal observations by state policy (Steinberg & Donaldson, 2016).

At the same time, findings from this study may not generalize to other states. Future work should examine the extent to which assimilation, and other sources of observer bias, exist in other field (i.e., non-experimental) settings. Recent research on the validity of observation scores in modern evaluation contexts tends to use experimental data from the MET project (e.g., Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). It remains to be seen if experimental findings generalize to non-experimental field settings.

Second, this study is limited in that it does not definitively identify the source of assimilation bias. The annual assignment of observations by state policy may drive assimilation bias. But it is possible that teacher assignment to a lower discrete LOE is a source of bias, despite the lack of evidence. Future research should disentangle these potential sources from one another. Qualitative or survey methods may be well-suited for such investigations. Additionally, the 2SLS RDD used in this study is unable to identify the specific mechanisms explaining assimilation bias. Qualitative methods may also be better suited in explorations about the extent to which observers consciously engage in assimilation bias and, if assimilation bias is conscious, why observers engage in this practice.

The third potential limitation is that moderation analyses using measures of administrator skill and effectiveness may not have detected heterogeneous effects because true observer skill and effectiveness only weakly influence the moderators. Indeed, the determinants of administrator LOE and average observation scores include sources that may only weakly relate to observer skill/ expertise. At face value, the two other administrator moderators, administrator observation scores concerning teacher professional learning and teacher evaluation, capture some degree of observer skill/ expertise. But, in practice, administrator observation scores may suffer from weak validity. Future work concerning assimilation bias could explore the extent to which heterogeneous effects exist across different moderators.

Implications

Many teachers work in states attaching high-stakes consequences to evaluation scores, and observation scores tend to receive the greatest weight in the typical multiple measure evaluation systems (American Institutes for Research, 2016; Cohen & Goldhaber, 2016). Additionally, several states assign observations on the basis of prior-year teacher performance (American Institutes for Research, 2016). In conjunction with the magnitude of assimilation bias, these conditions suggest efforts should be made to reduce the effects or causes of assimilation bias.

Education agencies may be able to mitigate the effects of bias in observation scores via regression adjustment. Indeed, the authors of some studies examining observer bias call for such adjustments (e.g., Campbell & Ronfeldt, 2018). Regression adjustment controls for sources of bias, comparing teachers encountering similar context independent and context dependent sources of bias. For example, observation scores could be adjusted by controlling for the number of observations assigned by state policy, characteristics of students taught, etc. If policymakers

move towards this approach it underscores the importance of this study, the need for additional work examining other sources of bias in observation scores, and research examining methods capable of substantially removing sources of bias in observation scores. Unless we can identify the sources of bias, and methods capable of adequately controlling for bias, regression adjusted observation scores will remain biased. At the same time, Milanowski (2016) argues regression adjusted observation scores may mask true differences in instructional quality and could make it more difficult to decide if disadvantaged students are assigned lower quality teachers.

Notwithstanding the controversy concerning regression adjusted observation scores, regression adjustment does not address the root causes of observer bias. To the extent post-observation feedback is influenced by observer bias, this also means leaving the sources of observer bias unaddressed could inhibit the effectiveness of observations as a tool for teacher development. Suppose an observer directs a teacher assigned more observations by state policy to engage in more professional development, not because the teacher's observed teaching was of low quality, but because of assimilation bias. Such misdirection misallocates professional development resources and teacher time. This implies that policymakers and education agencies might consider addressing the root causes of assimilation and other forms of observer bias.

One way to eliminate assimilation bias arising from the differentiated assignment of observations by state policy is for policymakers to assign all teachers the same number of observations. Indeed, more than 15 states do not differentiate the assignment of observations by any determinant (e.g., prior-year performance, teaching experience) (American Institutes for Research, 2016). Although this would eliminate assimilation bias arising from policy-assigned observations, it implies that there is no meaningful variation in teacher effectiveness. A premise of several teacher observation systems is that teacher effectiveness varies, and therefore less

effective teachers need relatively more observations to improve their effectiveness (Steinberg & Donaldson, 2016). Policymakers could recognize the importance of this premise by assigning all teachers the same number of observations as lower performing or early career teachers, but this would increase the administrative burdens of teacher evaluation, which school administrators report is already very time intensive (Kraft & Gilmour, 2016; Rigby, 2015).

A second way to address the cause of assimilation bias arising from the assignment of observations by state policy is through observer professional development. One might hope that education agencies could use available measures of observer skill/ effectiveness to target professional development to subgroups especially afflicted by assimilation bias. However, moderation analyses suggest that the degree of assimilation bias does not depend on the examined moderators, suggesting that education agencies may need to provide professional development to all observers. Many education agencies may already schedule opportunities to mitigate assimilation bias via observer professional development: observer (re)certification trainings. To the extent these trainings already discuss other forms of bias, observer (re)certification may be able to address assimilation bias. Although addressing assimilation bias through annual trainings may not substantially reduce the negative effects on observation scores, the magnitude of assimilation bias detected in this study, number of states policies assigning differentiated observations, and high-stakes consequences attached to evaluation ratings in many states, implies that even small reductions in assimilation bias may be worthwhile.

References

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

American Institutes for Research. (2016). Databases on State Teacher and Principal Evaluation Policies (STEP Database and SPEP Database). Retrieved June 7, 2019, from http://resource.tqsource.org/stateevaldb/Compare50States.aspx

Arvey, R. D., & Murphy, K. R. (1998). Performance Evaluation in Work Settings. *Annual Review of Psychology*, *49*, 141–168.

Bernardin, J. H., Thomason, S., Ronald Buckley, M., & Kane, J. S. (2016). Rater Rating-Level Bias and Accuracy in Performance Appraisals: The Impact OF Rater Personality, Performance Management Competence, and Rater Accountability. *Human Resource Management*, *55*(2), 321–340. https://doi.org/10.1002/hrm.21678

Borman, W. C. (1975). Effects of Instructions to Avoid Halo Error on Reliability and Validity of Performance Evaluation Ratings. *Journal of Applied Psychology*, *60*(5), 556–560. https://doi.org/10.1037/0021-9010.60.5.556

Campbell, S. L., & Ronfeldt, M. (2018). Observational Evaluation of Teachers: Measuring More Than We Bargained for? *American Educational Research Journal*, 000283121877621. https://doi.org/10.3102/0002831218776216

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater Calibration When Observational Assessment Occurs at Large Scale: Degree of Calibration and Characteristics of Raters Associated with Calibration. *Early Childhood Research Quarterly*, *27*(3), 529–542. https://doi.org/10.1016/j.ecresq.2011.12.006

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

Cohen, J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. *Educational Researcher*, *45*(6), 0013189X16659442-0013189X16659442. https://doi.org/10.3102/0013189X16659442

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

Decotiis, T., & Petit, A. (1978). The Performance Appraisal Process: A Model and Some Testable Propositions. *The Academy of Management Review*, *3*(3), 635. https://doi.org/10.2307/257552

Engelhard, G. (1994). Examining Rater Errors in the Assessment of Written Composition With a Many-Faceted Rasch Model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Georgia Department of Education. (2012). *Teacher Keys and Leader Keys Effective Systems*. Retrieved from http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/Pilot Report 12-13-2012 FINAL Clean.pdf

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings* (No. 8882021513; pp. 1–33).

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

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7–8), 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies*, *79*(3), 933–959. https://doi.org/10.1093/restud/rdr043

Jacob, B. A., & Walsh, E. (2011). What's in a rating? *Economics of Education Review*, *30*(3), 434–448. https://doi.org/10.1016/j.econedurev.2010.12.009

Kimball, S. M., & Milanowski, A. (2009). Examining Teacher Evaluation Validity and Leadership Decision Making Within a Standards-Based Evaluation System. *Educational Administration Quarterly*, *45*(1).

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

Ladd, H. F., & Sorensen, L. C. (2017). Returns to Teacher Experience: Student Achievement and Motivation in Middle School. *Education Finance and Policy*, *12*(2), 241–279. https://doi.org/10.1162/EDFP_a_00194

Ludwig, J., & Miller, D. L. (2007). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. *The Quarterly Journal of Economics*, *122*(1), 159–208.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Models for Teacher Accountability* (No. 0833035428; pp. 1–191). Retrieved from http://www.questia.com/PM.qst?a=o&se=gglsc&d=102693148

Mclntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of Performance Ratings as Affected by Rater Training and Perceived Purpose of Rating. *Journal of Applied Psychology*, *69*(1), 10.

Mihaly, K., & McCaffrey, D. F. (2014). Grade-Level Variation in Observational Measures of Teacher Effectiveness. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation Systems* (1st ed.). San Francisco, CA: Jossey-Bass.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A Composite Estimator of Effective Teaching* (pp. 1–51). Retrieved from http://www.nbexcellence.org/cms_files/resources/Jan 2013 A Composite Estimator of Effective Teaching Research Paper.pdf

Milanowski, A. (2016). Lower Practice Ratings for Teachers of Disadvantaged Students: Bias or Reflection of Reality? (Or Just Murky Waters?). *41st Annual Conference of the Association for Education Finance and Policy*, 1–28. https://doi.org/10.1017/CBO9781107415324.004

Papay, J. P., & Kraft, M. A. (2013). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Park, Y. S., Chen, J., & Holtzman, S. L. (2015). Evaluating Efforts to Minimize Rater Bias in Scoring Classroom Observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation Systems* (pp. 381–414). https://doi.org/10.1002/9781119210856.ch12

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

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214. https://doi.org/10.1162/qjec.2010.125.1.175

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the Ratings: Assessing the Psychometric Quality of Rating Data. *Psychological Bulletin*, *88*(2), 16.

SAS. (2015). *Technical Documentation for 2015 TVAAS Analyses 1.1*.

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., & Garrett, R. (2016). Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure? *Educational Evaluation and Policy Analysis*, *XX*(X), 0162373715616249-. https://doi.org/10.3102/0162373715616249

Sumer, H. C., & Knight, P. A. (1996). Assimilation and Contrast Effects in Performance Ratings: Effects of Rating the Previous Performance on Rating Subsequent Performance. *Journal of Applied Psychology*, *81*(4).

Tennessee Board of Education. *Teacher and Principal Evaluation Policy*. , 5.201 § (2013).

Tennessee Department of Education. (2016). *Evaluation | TEAM-TN*. Retrieved from http://team-tn.org/evaluation/

Tennessee General Assembly. *Tenure*. , (2016).

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and Rater Factors Affecting Rating Behavior. *Group & Organization Management*, *30*(1), 89–98. https://doi.org/10.1177/1059601104267920

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.

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology*, *95*(3), 546–561. https://doi.org/10.1037/a0018866

Wherry, R. J., & Bartlett, C. J. (1982). THE CONTROL OF BIAS IN RATINGS: A THEORY OF RATING. *Personnel Psychology*, *35*(3), 521–551. https://doi.org/10.1111/j.1744-6570.1982.tb02208.x

Notes

[1] Campbell and Ronfeldt (2018) are an exception, concluding that the influence of student and teacher race, and prior-year student achievement, on observation scores represent a form of observer bias.

[2] Prior work also examines nonsystematic (i.e. random) sources of rater bias (Mclntyre, Smith, & Hassett, 1984). A discussion of random error in ratings is beyond the scope of this paper.

[3] Others refer to "context-dependent bias" as "differential rater functioning" (e.g. Park, Chen, & Holtzman, 2015).

[4] TDOE does not collect performance measures for all teacher observers. However, more than 85 percent of all observers are school administrators.

[5] Student outcomes include two categories: "achievement" and "growth" scores. Achievement scores are district- or school-wide measures of student outcomes including graduation or attendance rates, or test-based outcomes such as ACT scores. The source of teacher growth scores depends on whether the teacher teaches a tested subject. Teachers of tested subjects receive a value-added score produced by the Tennessee Value-Added Assessment System (TVAAS). Teachers of untested subjects do not receive individual TVAAS scores. Growth scores for more than 80% of these latter teachers are based on a school-wide value-added score produced by TVAAS (for details see SAS, 2015). Growth scores for the remaining "untested" teachers are based on other value-added scores (e.g. subject-specific), portfolio scores (e.g. Fine Arts portfolios), or assessment scores (e.g. standardized K-2 student assessments).

[6] There is another discontinuity in policy-assigned observations at the LOE-cont 200 threshold. However, I do not discuss this threshold because less than one percent of all Tennessee teachers receive an LOE-cont below 200.

[7] Tenure status is determined by crossing an intermediate discrete LOE threshold (Tennessee General Assembly, 2016).

[8] Furthermore, neither school nor teacher fixed effects will address a plausible source of endogeneity. Observer discretion is plausibly informed by time-varying teacher motivation to improve and time-varying student behaviors (e.g., disruptive students), each of which likely affects observation scores. Because this source of endogeneity is time-varying within teacher, teacher fixed effects cannot remove it. Nonetheless, I explore the viability of teacher fixed effects, regressing the difference between total observations received and number assigned by state policy (i.e., "non-compliant" observations) on teacher, school x year, and year fixed effects. The fixed effects model finds that within-teacher variation accounts for approximately 30 percent of the variation in non-compliant observations. This means approximately 30 percent of the potentially endogenous variation in non-compliant observations remains in the fixed effect model.

Between-teacher differences in motivation to improve also cast serious doubt on the ability of school fixed effects to control for the endogenous receipt of observations.

[9] Previous research finds observers tend to generate more accurate scores in the environmental domain of teaching, relative to the instructional domain (Cash, Hamre, Pianta, & Myers, 2012).

[10] The Imbens-Kalyanaraman estimator identified an optimal bandwidth of 20, and Ludwig and Miller's cross-validation method (2007) produced an optimal bandwidth of 75. Considering that the difference between adjacent thresholds is 75, this bandwidth is unreasonably large.

[11] As discussed in footnote i, student outcomes are achievement or growth scores. Each score is an integer. The polychoric correlation between prior-year and contemporaneous achievement (growth) scores is 0.37 (0.50).

[12] Genuine within-year teacher growth could also explain positive within-year trends in observation scores. Indeed, observational processes aim to accomplish such growth. This is precisely why I use first scores as the dependent variable (see Methods section).

[13] Additionally, administrator observation scores are determined by survey-based input from school faculty/ staff concerning, in part, the professional growth of teachers. To the extent strategic low-to-high scoring of teachers influences faculty/ staff survey responses, administrators also have some control over this source of administrator observation scores.

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

[15] 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.

[16] Administrative data include observer identifiers, but these identifiers capture who enters teacher observation data into Tennessee's data management system. Because the person entering the data is not necessarily the person who conducted the observation, quartiles are based on school-level means.

Tables

Table 1

Descriptive Statistics

	Mean (SD)		
		RDD BW=40	
	Population	200-Threshold	425-Threshold
Observation Score	3.87	3.11	4.06
	(0.64)	(0.62)	(0.52)
Observations	2.37	3.48	1.84
Received	(1.25)	(1.28)	(0.96)
Yrs Exp	12.66	14.87	13.11
	(10.52)	(9.15)	(9.35)
BA+	61%	61%	58%
Female	79%	73%	80%
Nonwhite	14%	16%	5%
Prof Cert	78%	100%	88%
Discrete LOE1	0.44%	12.09%	
Discrete LOE2	8.40%	87.91%	
Discrete LOE4	32.73%		44.31%
Discrete LOE5	36.18%		55.69%

Note: Standard deviations in parentheses. Nonwhite is an indicator taking a value of one if the teacher is nonwhite, and value of zero if the teacher is white. BA+ is an indicator taking a value of one if the teacher earned more than a bachelor's degree.

Table 2

Balance Tests

	200-Threshold			425-Threshold		
	w=20	w=30	w=40	w=20	w=30	w=40
Yrs Exp: App				0.15	0.18	0.05
				(0.32)	(0.25)	(0.21)
Yrs Exp: Prof	0.39	-0.74	-0.32	-0.20	-0.04	0.15
	(1.91)	(1.58)	(1.37)	(0.38)	(0.31)	(0.28)
Female: App				-0.03	0.01	< 0.01
				(0.05)	(0.04)	(0.03)
Female: Prof	-0.03	-0.03	-0.04	0.02	0.01	0.01
	(0.10)	(0.09)	(0.08)	(0.02)	(0.01)	(0.01)
BA+: App				-0.06	-0.05	-0.02
				(0.05)	(0.04)	(0.04)
BA+: Prof	0.17	0.12	0.09	0.02	0.02	< 0.01
	(0.10)	(0.09)	(0.08)	(0.02)	(0.02)	(0.02)
Nonwhite: App				-0.02	-0.02	-0.02
				(0.02)	(0.02)	(0.02)
Nonwhite: Prof	-0.01	< 0.01	-0.02	-0.01	-0.01	< 0.01
	(0.09)	(0.07)	(0.06)	(0.01)	(0.01)	(0.01)
N(Tch-Yrs)	849	1434	2267	22607	33546	43893

Note: Estimates represent the total predicted change in the outcome. Standard errors in parentheses, clustered at the teacher level. Estimates at 200-threshold only use Professional teachers. OLS estimator employed to estimate all coefficients. BA+ is a binary variable indicating if a teacher reported having a degree higher than a BA/ BS. Nonwhite is an indicator signaling whether the teacher reported her ethnicity/ race as nonwhite or white. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3

Effects of Assigned Number of Observations on First Observation Scores

	200-Threshold			425-Threshold		
	w=20	w=30	w=40	w=20	w=30	w=40
Obs Freq	-0.30	-0.19	-0.22	-0.16**	-0.14**	-0.10**
	(0.23)	(0.20)	(0.17)	(0.06)	(0.05)	(0.04)
1st Stage: App				0.86***	0.84***	0.92***
				(0.13)	(0.11)	(0.09)
1st Stage: Prof	0.48**	0.42**	0.45***	0.21***	0.23***	0.26***
	(0.16)	(0.14)	(0.12)	(0.03)	(0.02)	(0.02)
1st-Stage: F-statistic	9.59**	9.76**	13.92***	44.54***	70.58***	118.03***
N(Tch-Yrs)	849	1434	2267	22607	33546	43893

Note: 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4

Effect of Crossing into Lower Discrete LOE

	275-Threshold			350-Threshold		
	w=20	w=30	w=40	w=20	w=30	w=40
Crossing to Lower LOE	-0.01	-0.01	> -0.01	0.05	0.06*	0.05*
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)
N(Tch-Yrs)	15699	23735	31568	9715	14497	18946

Note: Standard errors clustered at teacher level. The predictor of interest is crossing from discrete LOE4 to LOE3, or LOE3 to LOE2. Each model controls for teacher demographics, school-level teacher effectiveness, LOE-cont, month of first observation, and domains scored on first observation. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5

Sensitivity Tests: 425-Threshold Only

	w=20	w=30	w=40
Panel A: Low-to-High Scoring DV = First Score – Mean Score			
Obs Freq	< 0.01	-0.02	-0.01
	(0.03)	(0.02)	(0.01)
N(Tch-Yrs)	22607	33546	43893
Panel B: Easier Scoring DV = First Score			
Obs Freq	-0.17	-0.14	-0.07
	(0.19)	(0.13)	(0.09)
N(Tch-Yrs)	13248	19608	25545
Panel C: Switching Tested Status DV = Change in Tested Status			
App	-0.04	-0.01	> -0.01
	[0.04]	[0.03]	[0.03]
Prof	< 0.01	-0.01	-0.02
	[0.02]	[0.01]	[0.01]
	22607	33546	43893

Note: Standard errors clustered at teacher level. Samples in Panel A restricted to teachers receiving more than one observation per year. The outcome in Panel B is the difference between first scores and mean TEAM observation scores. Models in Panel A and B control for teacher demographics, school-level teacher effectiveness, LOE-cont, month of first observation, and domains scored on first observation. The model in Panel C drops controls for month of first observation and domains scored. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6

Effects of Crossing 425-Threshold on Teacher Improvement Efforts

		$w = 20$	$w = 30$	$w = 40$
Sum: Svy Hrs in PD (*PDhrs*)	App	-1.57	0.2	-0.01
		[1.39]	[1.29]	[1.19]
	Prof	-0.21	-0.17	-0.47
		[0.66]	[0.54]	[0.45]
	F-statistic	0.69	0.06	0.53
		(0.50)	(0.94)	(0.59)
	N(Tch-Yrs)	25486	34787	40883
Sum: Svy Exerted More Effort (*effortsum*)	App	0.05	-0.11	-0.13
		[0.17]	[0.14]	[0.12]
	Prof	-0.12*	-0.03	-0.03
		[0.06]	[0.05]	[0.04]
	F-statistic	2.02	0.44	0.78
		(0.13)	(0.65)	(0.46)
	N(Tch-Yrs)	20711	29329	35173
Sum: Svy Hrs Improved Instruction (*insthrs*)	App	-0.91	0.86	-4.25
		[10.99]	[8.62]	[7.42]
	Prof	-1.7	-2.64	-3.58
		[3.49]	[2.86]	[2.49]
	F-statistic	0.12	0.43	1.20
		(0.88)	(0.65)	(0.30)
	N(Tch-Yrs)	3471	5390	7052

Note: Teacher clustered standard errors in brackets, *p*-values 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. * $p < 0.05$.

Table 7

Moderation by Measures of Administrator Effectiveness and Performance

	w=20	w=30	w=40
Panel A: Administrator LOE-Cont			
Obs Freq: 1st Quartile	-0.13***	-0.14***	-0.14***
	[-0.19, -0.07]	[-0.20, -0.09]	[-0.19, -0.09]
Obs Freq: 2nd Quartile	-0.16***	-0.15***	-0.13***
	[-0.21, -0.11]	[-0.20, -0.11]	[-0.16, -0.09]
Obs Freq: 3rd Quartile	-0.13***	-0.13***	-0.11***
	[-0.18, -0.08]	[-0.17, -0.09]	[-0.14, -0.08]
Obs Freq: 4th Quartile	-0.11***	-0.12***	-0.10***
	[-0.16, -0.06]	[-0.16, -0.08]	[-0.14, -0.07]
N(Tch-Yrs)	25915	38454	50526
Panel B: Administrator Observation Scores			
Obs Freq: 1st Quartile	-0.14***	-0.15***	-0.13***
	[-0.20, -0.09]	[-0.20, -0.10]	[-0.17, -0.09]
Obs Freq: 2nd Quartile	-0.13***	-0.14***	-0.12***
	[-0.18, -0.09]	[-0.18, -0.10]	[-0.16, -0.09]
Obs Freq: 3rd Quartile	-0.11***	-0.13***	-0.11***
	[-0.16, -0.07]	[-0.17, -0.09]	[-0.14, -0.08]
Obs Freq: 4th Quartile	-0.13***	-0.14***	-0.12***
	[-0.18, -0.08]	[-0.18, -0.09]	[-0.15, -0.08]
N(Tch-Yrs)	25663	38077	50050

Note: 95% confidence intervals in brackets. Standard errors clustered at teacher level. Controls include teacher demographics, school-level teacher effectiveness, LOE-cont, month of first observation, and domains scored on first observation. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 8

Moderation by Measures of Administrator Skill

	w=20	w=30	w=40
Panel A: Skills as Teacher Evaluator			
Obs Freq: 1st Quartile	-0.14***	-0.15***	-0.13***
	[-0.20, -0.09]	[-0.20, -0.10]	[-0.17, -0.09]
Obs Freq: 2nd Quartile	-0.13***	-0.14***	-0.12***
	[-0.18, -0.09]	[-0.18, -0.10]	[-0.16, -0.09]
Obs Freq: 3rd Quartile	-0.11***	-0.13***	-0.11***
	[-0.16, -0.07]	[-0.17, -0.09]	[-0.14, -0.08]
Obs Freq: 4th Quartile	-0.13***	-0.14***	-0.12***
	[-0.18, -0.08]	[-0.18, -0.09]	[-0.15, -0.08]
N(Tch-Yrs)	25663	38077	50050
Panel B: Skills in Supporting Teacher Professional Learning			
Obs Freq: 1st Quartile	-0.13***	-0.15***	-0.14***
	[-0.18, -0.08]	[-0.19, -0.11]	[-0.17, -0.10]
Obs Freq: 2nd Quartile	-0.09***	-0.10***	-0.09***
	[-0.14, -0.04]	[-0.15, -0.06]	[-0.13, -0.06]
Obs Freq: 3rd Quartile	-0.11***	-0.13***	-0.11***
	[-0.16, -0.07]	[-0.17, -0.09]	[-0.14, -0.08]
Obs Freq: 4th Quartile	-0.10***	-0.11***	-0.10***
	[-0.15, -0.05]	[-0.16, -0.07]	[-0.14, -0.06]
N(Tch-Yrs)	23523	35075	46292

Note: 95% confidence intervals in brackets. Standard errors clustered at teacher level. Controls include teacher demographics, school-level teacher effectiveness, LOE-cont, month of first observation, and domains scored on first observation. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 9

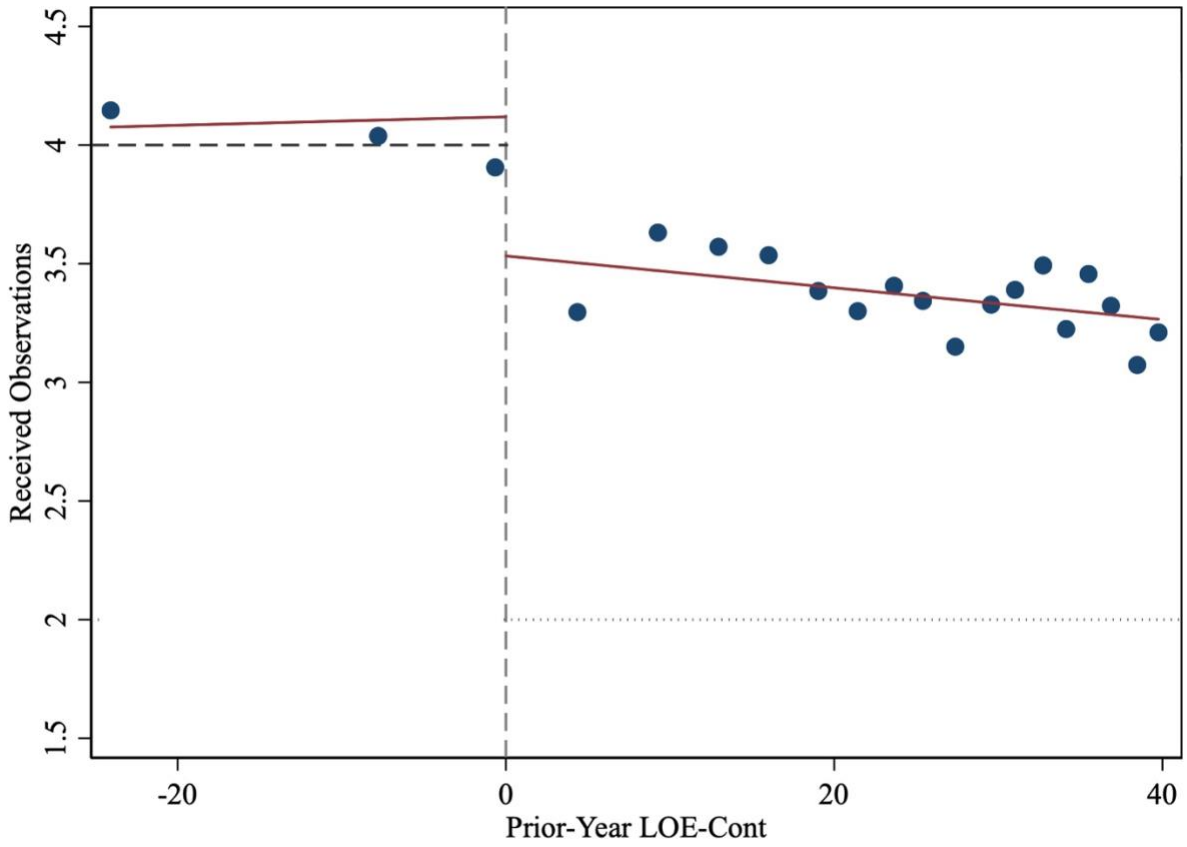
Moderation by Teacher Years of Experience

	w=20	w=30	w=40
Panel A: Continuous Measure			
Obs Frequency	-0.16**	-0.14***	-0.09**
	[0.05]	[0.04]	[0.03]
Interaction	< 0.01	< 0.01	< 0.01
	[< 0.01]	[< 0.01]	[< 0.01]
Panel B: Binned Measures			
Obs Frequency	-0.20***	-0.18***	-0.14***
	[0.04]	[0.04]	[0.03]
Int: Yrs Exp 3 through 5	-0.02	-0.01	-0.02
	[0.03]	[0.03]	[0.02]
Int: Yrs Exp 6 or More	-0.01	< 0.01	< 0.01
	[0.04]	[0.03]	[0.03]
N(Tch-Yrs)	22607	33546	43893

Note: Panel A interacts a continuous measure of teacher years of experience with the frequency of observations. Panel B interacts a binned measure of teacher years of experience with the frequency of observations and replaces the continuous years of experience control with the binned measure. Standard errors clustered at teacher level. Controls include teacher demographics, school-level teacher effectiveness, LOE-cont, month of first observation, and domains scored on first observation. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

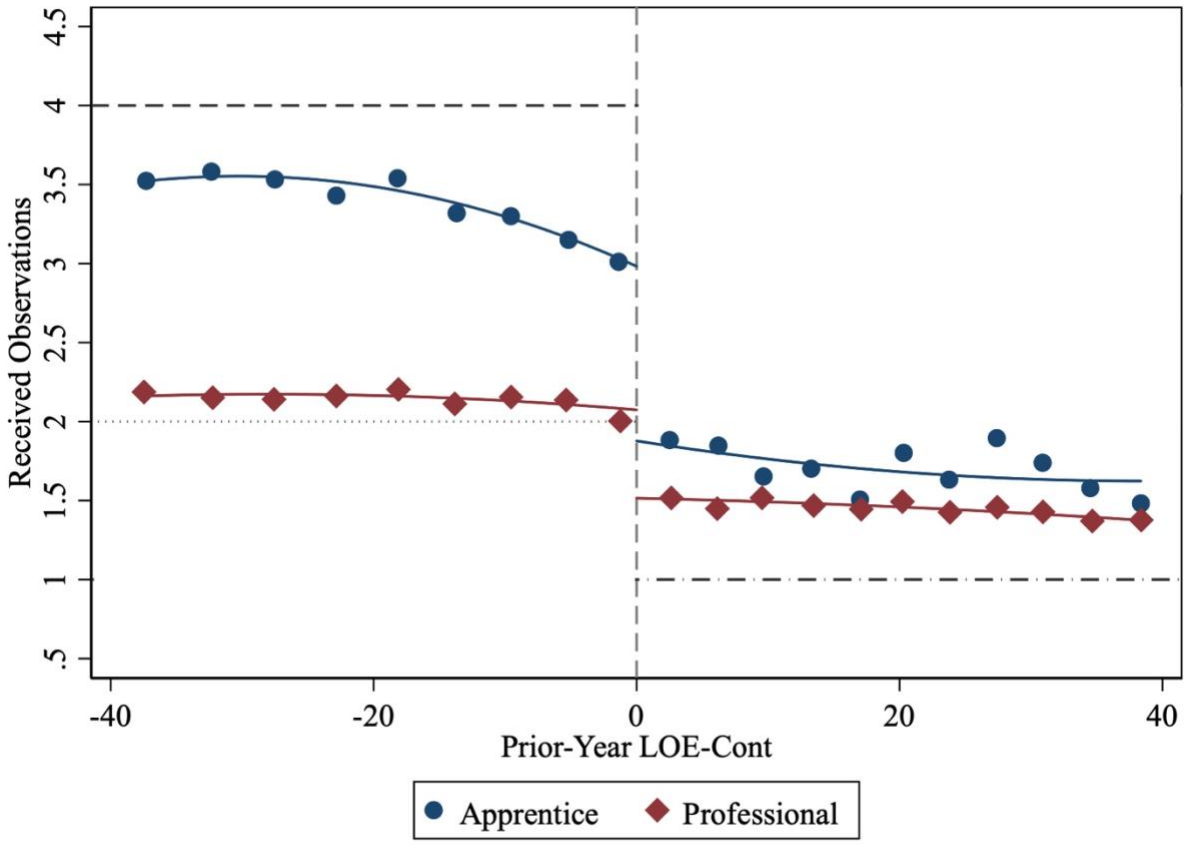
Figures

Figure 1. Binned Scatterplot: Observations vs Prior-Year LOE-Cont at 200-Threshold



Note: Plotted points are the mean number of observations received within bins of four. A discontinuity in the number of policy-assigned observations exists at LOE-cont = 200. Horizontal dashed lines represent the number of observations assigned to Professional teachers by state policy. Policy assigns Professional teachers: above 200 two observations, and below 200 four.

Figure 2. Binned Scatterplot: Observations vs Prior-Year LOE-Cont at 425-Threshold



Note: Plotted points are the mean number of observations received within bins of four. 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. Curves are second-order polynomials.

Figure 3. Histogram of Prior-Year LOE-Cont

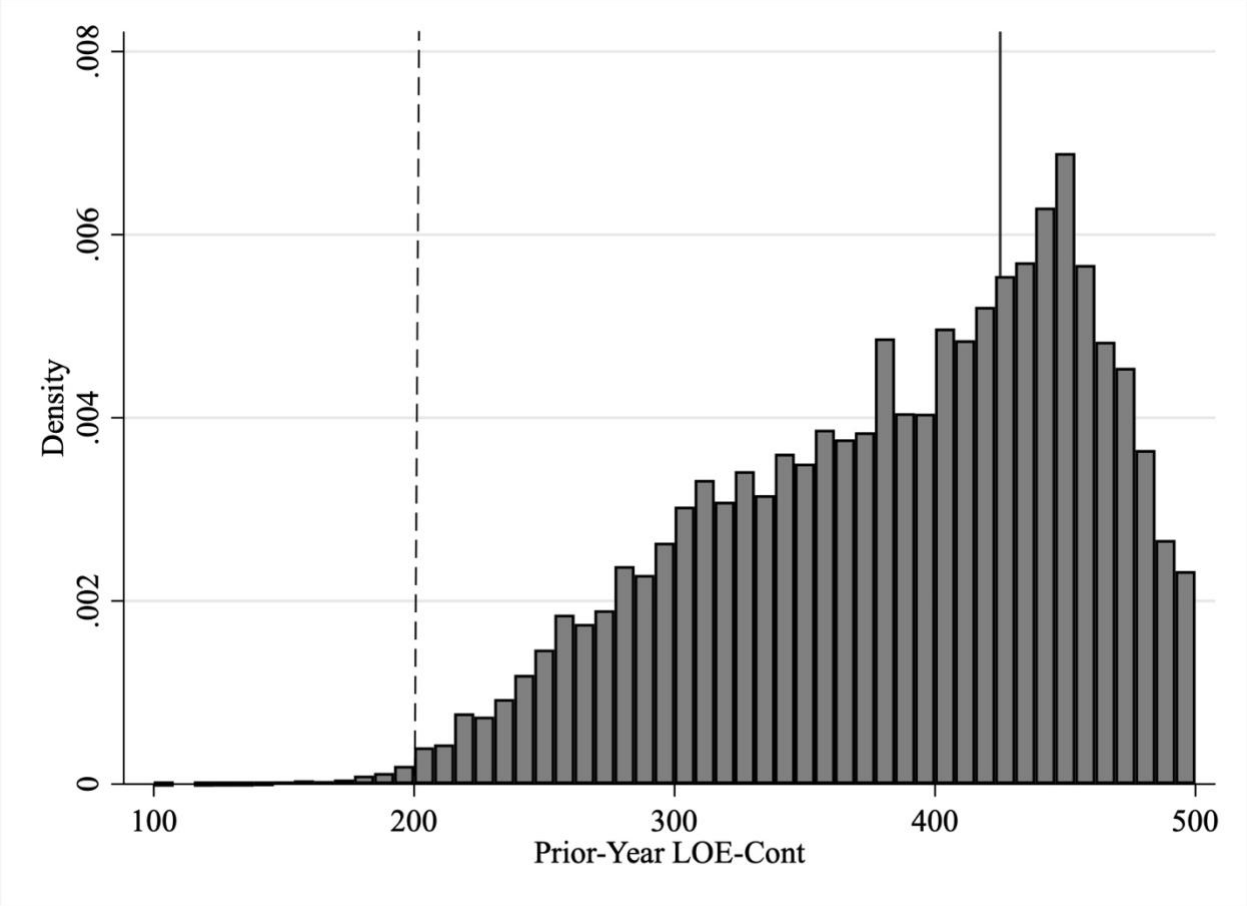


Figure 4. Histogram of Prior-Year LOE-Cont Transformed by Modulus 5