AERA Open
April-June 2020, Vol. 6, No. 2, pp. 1–16
DOI: 10.1177/2332858420929276
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2020. http://journals.sagepub.com/home/ero

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

Seth B. Hunter
George Mason University

Several state policies link high-stakes consequences to teacher evaluations, which tend to be heavily weighted by observation scores. However, research has only recently investigated the validity of these scores in field settings. This study examines the sensitivity of teacher observation scores to the number of observations assigned by state policy and the assignment of prior-year composite measures produced by the evaluation system. Regression discontinuity and local regression designs exploit discontinuities in both assignment processes. The evidence suggests that assignment to a lower prior-year composite score does not bias observation scores, but the assignment to more policy-assigned observations introduces substantial negative bias. The degree of negative bias is most pronounced among early-career teachers, as suggested by theory. Implications are discussed.

Keywords: *educational policy, evaluation, regression analyses, regression discontinuity, school/teacher effectiveness, validity/reliability*

Throughout the first two decades of the 21st century, many local and state education agencies in the United States substantially reformed teacher evaluation (Putman et al., 2018; Steinberg & Donaldson, 2016; Walsh et al., 2017). One of the most widely adopted reforms substantially changed teacher observation systems (Cohen & Goldhaber, 2016). Many education agencies altered teacher observation systems by increasing the frequency of teacher observations and adopting standards-based protocols (e.g., the widely adopted Framework for Teaching; American Institutes for Research, 2016; National Council on Teacher Quality, 2019a, 2019b). Additionally, some state education agencies began assigning observations based on combinations of teacher prior-year *performance* (i.e., how teachers taught) or *productivity* (i.e., changes in schooling outcomes such as growth in student achievement scores; American Institutes for Research, 2016; National Council on Teacher Quality, 2019a; Putman et al., 2018; Walsh et al., 2017). Historically, states differentiating the assignment of observations did so based on teaching experience or tenure, not prior-year performance or productivity (Steinberg & Donaldson, 2016). Some scholars refer to systems incorporating these reforms as "next-generation" evaluation systems (Campbell & Ronfeldt, 2018; Steinberg & Donaldson, 2016).

Research finds that observation scores (i.e., a performance measure) tend to receive the most weight among the performance and productivity measures informing high-stakes outcomes (Cohen & Goldhaber, 2016; Steinberg & Donaldson, 2016). By the mid–2010s, nearly half of all states attached punitive consequences (e.g., loss of tenure) to low teacher performance or productivity (American Institutes for Research, 2016). The weight states typically gave to observation scores means these scores play a significant role in the allocation of high-stakes consequences. Despite the importance of these observation scores, only a few studies examine the validity of teacher observation scores in next-generation systems, and many of these studies use data from the well-known Measures of Effective Teaching study (e.g., Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). Stated differently, we are learning a great deal about the validity of researcher-generated observation scores produced in one small-scale experimental setting, but there is still a great deal to learn about the properties of observation scores generated by practitioners (i.e., typically school administrators) in field settings.

Understanding the properties of observation scores in field settings is critical because bias in observation scores (i.e., deviations from "true" scores based solely on teacher performance) may undermine the primary goals of teacher evaluation. Although teacher evaluation systems inform personnel decision making, state education agency leaders tend to emphasize teacher evaluation as developmental (Almy, 2011; Donaldson & Papay, 2014; Georgia Department of Education, 2012; Tennessee Department of Education, 2016). However, scholars argue that the effectiveness of evaluation as a developmental tool depends on employee trust in the evaluation system (Lane, 2019; K. R. Murphy & Cleveland, 1995). Indeed, Donaldson (2012) finds that teacher trust in evaluation systems diminishes when teachers believe their observation scores are influenced by external factors.

Previous research concerning bias in next-generation observation scores primarily focuses on 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 race (Campbell & Ronfeldt, 2018), teaching experience (Jacob & Walsh, 2011), and the prior-year achievement of incoming students (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). However, as the authors of these previous studies discuss, these effects or relationships may not represent bias.[1]

This study extends the literature concerning theories of bias and research on the validity of next-generation observation scores by exploring two previously unexamined sources of bias: prior-year teacher *effectiveness* scores (i.e., a composite measure of teacher performance and productivity) and the number of observations assigned to teachers by state policy. After the start of each school year, teachers in the study setting receive their effectiveness score that determines how many observations they should receive per state policy. Theories of bias in observation scores imply that observers might (un)consciously base subsequent observation scores on prior-year effectiveness scores and the number of observations assigned to a teacher independent of the teacher's observed performance. Psychologists label this "assimilation bias," because observers generate scores assimilating toward their expectations of employee performance (Sumer & Knight, 1996).

Conceptual Framework

Scholars define observer bias as the extent to which scores systematically[2] deviate from an employee's "true performance" (Bernardin et al., 2016; Park et al., 2015; Wherry & Bartlett, 1982). Few studies examine observer bias in next-generation teacher observation systems, and even fewer use data from field settings (see below). Studies examining bias in next-generation systems tend to focus on the influence of "classroom characteristics" and teacher and observer demographics (e.g., Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). However, earlier psychological works recognizes other sources of bias. Indeed, some work implies that bias may arise from features of the evaluation system.

In broad terms, previous work explores *context-independent* or *context-dependent* bias.[3] Context-independent bias originates from the observer herself (Park et al., 2015). For example, observers who systematically issue lower ratings relative to other observers exhibit "severity" bias (Engelhard, 1994). 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 conditions including grade taught and student characteristics (Campbell & Ronfeldt, 2018; Graham et al., 2012; Mihaly & McCaffrey, 2014; Steinberg & Garrett, 2016).

More recent work examines the influence of classroom characteristics. Steinberg and Garrett (2016) and Campbell and Ronfeldt (2018) use Measures of Effective Teaching (MET) experimental data to identify the extent to which *researcher*-generated observation scores are influenced by incoming student: achievement scores, race/ ethnicity, and gender. Steinberg and Garrett (2016) find that incoming achievement scores positively influence observational ratings and conclude that the relationship may capture observer bias or genuine teacher responses to student instructional needs. 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; these authors imply that classroom characteristics are a source of observer bias.

Previous work also suggests that employee (i.e., teacher) gender, race, and observer-employee race congruence influenced performance ratings. Using MET data, researchers find that male teachers receive systematically lower observation scores than females (Campbell & Ronfeldt, 2018). Research within and beyond educational settings finds that Black employees (i.e., teachers) tend to receive lower ratings than Whites (Arvey & Murphy, 1998; Campbell & Ronfeldt, 2018), and employees sharing the same race as their observer (i.e., race-congruent) tend to receive higher observation scores (Arvey & Murphy, 1998).

Assimilation Bias

Psychologists also argue that information about prior performance or productivity can introduce context-dependent bias by influencing observer expectations of employee performance independent of subsequently observed performance (Hogan, 1987; Kingstrom & Mainstone, 1985; Lawler, 1967; Wherry & Bartlett, 1982). Psychologists label bias toward information about the employee's prior performance or productivity *assimilation bias* (Sumer & Knight, 1996; Wang et al., 2010). The theory of assimilation bias does not suggest that information about prior performance or productivity should be unrelated to subsequent performance scores. Indeed, researchers consider measures of teacher performance and productivity unreliable if past and contemporary scores are not strongly related (e.g., Amrein-Beardsley, 2008; Brennan, 2001; Grossman et al., 2013; Hill et al., 2012; Ho & Kane, 2013). Instead, assimilation bias arises when information "shocks" observer expectations of employee performance, independent of observed performance. Importantly, assimilation bias does not depend on an individual's prior performance relative to her peers' prior performances, but on the individual's prior performance relative to a performance scale. For example, negative assimilation bias will theoretically still affect teachers in a school filled with only teachers who received the highest prior-year performance rating on a performance scale.

It is difficult to disentangle assimilation bias from genuine differences in employee performance, especially in field settings; however, laboratory research finds evidence of assimilation bias (Sumer & Knight, 1996; Wang et al., 2010). Laboratory studies have assigned participants vignettes about a hypothetical employee's performance and asked participants to rate the performance. Before reading the vignettes, participants were randomly assigned an additional vignette about the hypothetical employee's prior performance. Some of the second vignettes described high prior performance, others described low prior performance. Each study found some evidence that participant-generated ratings positively correlated with the prior performance described in the randomly assigned vignette (Sumer & Knight, 1996; Wang et al., 2010).

Moderator: Job Experience

Shocks to observer expectations about employee performance theoretically depend on the amount of information an observer has about employee prior performance (Hogan, 1987; Lawler, 1967; Wherry & Bartlett, 1982). Indeed, psychologists argue that assimilation bias may be more pronounced among newer employees (Hogan, 1987; Lawler, 1967; K. R. Murphy & Deshon, 2000; Wherry & Bartlett, 1982). As employees accrue performance histories relevant performance behaviors become accessible to the observer and will theoretically counter biasing shocks to observer expectations (Conway, 1996; Kingstrom & Mainstone, 1985).

Hypotheses

This study hypothesizes that prior-year effectiveness scores and the assignment of some number of observations by state policy are sources of assimilation bias. As discussed above, previous laboratory studies suggest that observer knowledge about employee prior performance may be a source of assimilation bias. Although few, if any, studies examine the assignment of observations as a source of bias, it may shock observer expectations about employee performance. Observers may (un)consciously believe that employees assigned more observations are less effective, independent of subsequently observed performance. Otherwise, why would the teacher evaluation system assign more observations? Indeed, in the study context, some teachers are assigned more observations due to lower prior-year effectiveness scores. However, the relationship between prior-year effectiveness and observation assignment is not perfectly collinear for all teachers in the study context, allowing for separate examinations of these potential sources of assimilation bias. Additionally, based on previous research concerning the theoretical moderation of assimilation bias, this study also hypothesizes that the degree of bias will be the strongest for early-career teachers.

Study Context: Teacher Evaluation and Observation in Tennessee

In the early 2010s, the Tennessee Department of Education (TDOE) adopted the Tennessee Educator Acceleration Model (TEAM) teacher evaluation system. The TEAM system includes several reforms, but two of the most prominent are the introduction of a composite measure of teacher effectiveness and changes to teacher observation policy. TDOE labels the composite measure the "Level of Effectiveness" (LOE). LOE and *certification status*, which is effectively determined by years of experience, became the determinants of policy-assigned observations.

Level of Effectiveness Scores

TDOE generates two expressions of teacher LOE scores. *LOE-cont*, the first expression, is a continuous, composite measure determined by observation scores and student outcomes. Student outcomes include two categories: "achievement" and "growth" scores. Achievement scores are district- or school-wide student outcomes, including graduation rates and test scores. The source of teacher growth scores depends on whether the teacher taught a tested subject. Teachers of tested subjects receive a value-added score produced by the Tennessee Value-Added Assessment System (i.e., the well-known TVAAS score). 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 (e.g., subject-specific), portfolio (e.g., Fine Arts portfolios), or assessment scores (e.g., standardized K–2 student assessments), chosen by the teacher and her evaluator at the beginning of the school year prior to the teacher's first observation. At least 50% of LOE-Cont for teachers of tested subjects is based on student outcomes with the remainder determined by observation scores. For teachers of untested subjects, at least 40% of LOE-Cont is determined by student outcomes with the rest based on observation scores. TDOE converts LOE-Cont to the second expression of effectiveness scores, *discrete LOE*. LOE-Cont scores within [100, 200), [200, 275), [275,350), [350, 425), or [425, 500] are respectively assigned discrete LOE of LOE1, LOE2, LOE3, LOE4, or LOE5.

After the start of each school year during the study period (2012–2013 through 2014–2015), each teacher received her own discrete LOE score while school administrators, and the observers who were not school administrators, received the discrete LOE of teachers they were to evaluate. It was important for observers to know a teacher's discrete LOE because it determined the number of observations assigned by state policy. During the study period, the state information management system did not tell observers how many times they should observe a teacher; instead, observers were expected

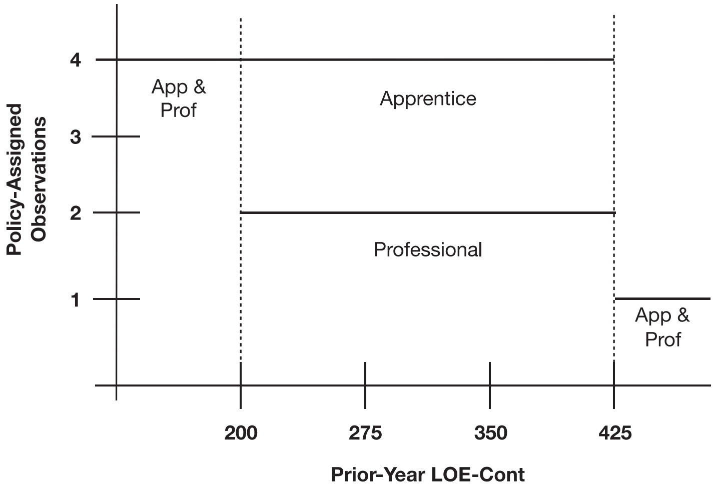

FIGURE 1. *Tennessee Board of Education observation assignment policy.*

to determine the number of observations needed based on discrete LOE and certification status. Supplementary Appendix A presents strong evidence that observers issued observations based on these two policy-defined determinants. Additionally, no observers or teachers ever received LOE-Cont; thus, observers nor teachers knew if a teacher assigned to a discrete LOE had a "high" or "low" LOE-cont within the discrete LOE. Prior-year discrete LOE is used to test the hypotheses that prior-year effectiveness scores introduce assimilation bias.

Tenure. Teachers with at least 5 years' experience who have not received an LOE-Cont[4] below 350 in the past 3 years are eligible for "tenure" (Tennessee General Assembly, 2016); however, tenure protections in Tennessee are weak. A tenured Tennessee teacher can lose tenure due to 2 years of low effectiveness scores (Tennessee General Assembly, 2016).

TEAM Observation Policy

The Tennessee State Board of Education (2013) adopted several observation policies in the early 2010s. First, only annually certified observers could conduct formal observations (henceforth "observations"). Observer certification focuses on generating accurate and reliable observation scores and facilitating pre- and postobservation conferences to improve teacher performance (Alexander, 2016). The majority of observers[5] are principals or assistant principals, a small percentage of observers are full-time teacher evaluators, and the remainder are central office personnel (Alexander, 2016).

Second, TEAM observers must use the TEAM standards-based protocol, or "TEAM rubric" (see Supplementary Appendix B). The TEAM rubric is used to measure teacher performance concerning Planning, Instruction, and the Classroom Environment. Third, after the start of each school year, TEAM teachers receive their discrete LOE and certification status, which determines the number of observations assigned to teachers by state policy. Certification status indicates whether a teacher has taught fewer than 4 years ("Apprentice"), or not ("Professional").

Policy assigns teachers with a prior-year LOE5 (LOE-cont $\geq$ 425) one observation, and teachers with a prior-year LOE1 (LOE-cont $<$ 200) four. The number of observations assigned to teachers with a prior-year LOE2 through LOE4 (200 $\leq$ LOE-cont $<$ 425) is two or four, depending on certification status. Thus, there are two discontinuities in the number of policy-assigned observations at the LOE-cont 425 threshold, and one at the 200 threshold. The only discrete LOE within which there are discontinuities in observations are LOE2 through LOE4. See Figure 1 for a graphical representation of the observation assignment policy. According to state policy, districts or schools could exercise discretion and conduct more than the policy-assigned number of observations, though no teacher should receive less (Tennessee State Board of Education, 2013). Survey data collected by the Tennessee Department of Education suggests that each observation lasts about 30 minutes[6] (Periscopic, 2019).

Classroom Observations and the Improvement of Teacher Performance

The TEAM theory of action asserts that observations will improve teacher performance as measured by the TEAM rubric (Alexander, 2016; TDOE, 2016), which describes a range of standards-based teacher behaviors, from *unsatisfactory* (1) to *exemplary behavior* (5) (Daley & Kim, 2010). After an observation, the observer is expected to hold a postobservation conference and supply feedback aligned to the TEAM rubric. TDOE expects observers to work with teachers to develop improvement plans (e.g., coaching, self-study) as needed.

Data

This study uses teacher panel data linked to more than 80% of Tennessee school districts from 2012–2013 through 2014–2015. The data includes teacher gender, level of education, race, years of experience, observation scores, and prior-year measures of discrete LOE and LOE-Cont. Observation records are at the observation-occurrence level and include observation dates and scores.

Some robustness analyses use data from the Tennessee Educator Survey (TES). Each spring, all Tennessee teachers receive the TES. Response rates exceeded 50% during the study period.

Method

This study explores if assignment to a lower discrete LOE and the assignment of more observations introduces assimilation bias in teacher observation scores. There are several opportunities for exploring these two potential sources of bias.

Three opportunities assign teachers more observations *and* to a lower discrete LOE (see Figure 1). The first compares Professional teachers just below the prior-year LOE-cont 200 threshold to Professionals just above, which also compares teachers assigned four observations to teachers assigned two (see Figure 1). The second and third opportunities compare teachers just above and below the 425 threshold and contrasts Apprentices assigned four observations to those assigned one (second opportunity) and Professionals assigned two observations instead of one (third opportunity). These three cases alone cannot definitively disentangle the influence of assigning teachers observations from assigning teachers to a lower LOE. However, there is a discontinuity in assigned observations that is not entangled with a change in discrete LOE.

Professional teachers with a prior-year discrete LOE2, LOE3, or LOE4 are assigned two observations, while Apprentices sharing the same prior-year discrete LOE are assigned four (see Figure 1). Comparing all Professional teachers to all Apprentice teachers in this range of prior-year LOE effectively contrasts mid- and late-career teachers to early-career teachers. Previous research finds that more experienced teachers tend to perform higher than less experienced teachers, plausibly confounding returns to teacher experience with the assignment of fewer observations (Harris & Sass, 2011; Ladd & Sorensen, 2017; Papay & Kraft, 2013). I mitigate some of these potential confounding influences by only comparing Professional teachers with four years of experience to Apprentice teachers with 3 years.

The final three opportunities for comparison contrast teachers just below a prior-year LOE-Cont threshold to those just above, at thresholds where there are no discontinuities in assigned observations. Apprentices falling just to either side of the LOE-Cont 200 are assigned to different discrete LOE, but both groups are assigned four observations (see Figure 1). Similarly, Apprentices falling just to either side of the LOE-Cont 275 or 350 thresholds are assigned the same number of observations; the same is true for Professionals surrounding these two thresholds (see Figure 1).

Outcome Variable

A naïve analytical strategy might treat the average observation score of teacher i in year t (S_{it}) as the outcome, regressing S_{it} on the number of observations assigned to teacher i in year t using ordinary least squares. However, this strategy is problematic because the TEAM system aimed to improve S_{it}. If the theory of action holds, estimates of the relationship between S_{it} and assigned observations may capture assimilation bias and genuine teacher improvements brought about by observation processes. However, classroom observations cannot influence observation scores generated before the receipt of post-observation feedback. Thus, none of the observations received in year t can genuinely affect a teacher's first observation score $\left(S_{1it}\right)$. The naïve strategy is improved on by replacing S_{it} with S_{1it}, which has a mean and standard deviation of 3.87 and 0.64, respectively.

Regressions

Regression Discontinuity Designs. Crossing from just above to just below the prior-year LOE-Cont thresholds of 200, 275, 350, and 425 always assigns teachers to a lower discrete LOE (recall, educators never received LOE-Cont), and sometimes assigns more observations. This type of assignment process is well-suited for a regression discontinuity research design (RDD), represented by the following model:

$$S_{1it} = \delta\rho_{it} + \boldsymbol{A}h(\bullet) + \boldsymbol{BX}_{it} + \gamma_t + \omega_{it}, \left|LOECont_{it}\right| \leq w \tag{1}$$

where S_{1it} is the first score received by teacher i in year t, ρ_{it} a vector of two indicators signaling whether an Apprentice or Professional teacher is above or below a specified prior-year LOE-Cont threshold, and h is a second-order polynomial of LOE-Cont interacted with ρ_{it}, allowing the relationship between h and S_{1it} to vary across each threshold. Only teachers with a first score in year t and LOE-Cont score in year t – 1 are included in the analytical sample. Critically, h is the sole determinant of teacher discrete LOE, and ρ_{it} and h are the only determinants of the number of observations assigned to a teacher by state policy.

$\boldsymbol{X}_{it}$ is a vector of covariates, including teacher race/ ethnicity, gender, years of teaching experience, and level of education, which are included to increase precision. $\boldsymbol{X}_{it}$ also includes the month of the first observation and domains rated on the first observation. It is plausible that the timing of observations correlates with teacher performance (e.g., observers might have wanted to postpone difficult observations). Additionally, observers may score one domain more harshly than another[7]; γ_t is a year fixed effect to account for secular trends, and ω_{it} an idiosyncratic error term; *w* represents bandwidths of 20, 30, and 40 on either side of each threshold and brackets the Imbens and Kalyanaraman (2012) optimal bandwidth.[8] Standard errors are clustered at the teacher level.[9]

The relationship of interest (δ) captures two estimates, each of which is assumed to be the same across all discontinuities. The first relationship is based on a comparison of Apprentice teachers just below an LOE-Cont threshold to Apprentices just above, within bandwidths *w*. The second estimate is based on a similar comparison of Professionals. Thus, δ represents the effect of assigning teachers to a lower discrete LOE or more observations, moderated by career stage. RDDs assume that teachers are just to either side of each threshold according to a "locally random" process. If RDD assumptions are met, the only meaningful difference between the groups of teachers just to either side of a threshold is that the prior-year LOE-Cont score of the group just below the threshold "forces" that group into a lower discrete LOE or assigns them more observations. Therefore, the RDDs isolate the shock of assigning a teacher to a lower prior-year discrete LOE or more observations, overcoming limitations in previous field research about assimilation bias.

Local Regressions. I apply Equation 2 to explore the discontinuity in assigned observations across the Apprentice-Professional boundary within discrete LOE2 – LOE4.

$$\begin{aligned} S_{1it} &= \beta c_{it} + \boldsymbol{A}h(\bullet) + \boldsymbol{BX}_{it} + \boldsymbol{DL}_{it} \\ &+ \gamma_t + e_{it}, experience_{it} = \{3,4\} \end{aligned} \tag{2}$$

where all variables represent the same quantities as in Equation 1, c_{it} is a binary variable taking a value of one if teacher *i* was an Apprentice in year *t*, $\boldsymbol{L}_{it}$ is a vector of dummy variables indicating if the teacher is assigned to a prior-year discrete LOE1, LOE2, LOE3, LOE4, or LOE5, and e_{it} the error term. By including $\boldsymbol{L}_{it}$, Equation 2 compares Apprentice and Professional teachers assigned to the same prior-year discrete LOE. Importantly, Equation 2 only uses Apprentice teachers with 3 years of experience and Professional teachers with four years of experience; β captures the association between assignment to more observations (i.e., four instead of two) and first scores.

Validating Regression Discontinuity Designs

Manipulation of the Running Variable

An RDD is invalid if there is evidence of nonrandom assignment to either side of a threshold. Nonrandom assignment occurs when the forcing variable (i.e., prior-year LOE-Cont) is purposefully manipulated. The only individuals who can manipulate LOE-Cont scores are observers. One may be concerned that observers nonrandomly place teachers to either side of an LOE-Cont threshold for reasons related to teacher performance. For example, an observer may try to place a teacher above the 425 threshold because she believes the teacher is on a path toward improvement and does not need more policy-assigned observations.

However, the 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 determinants of LOE-Cont to manipulate LOE-Cont via observation scores. Observers could turn to historic student outcomes to predict contemporaneous scores. However, the correlations between prior-year and contemporaneous student-based outcomes[10] are below 0.50. Despite the implausibility of manipulation, I devise and apply a statistical test for manipulation under the assumption of observer prescience (Supplementary Appendix C discusses further the motivation for this test). If there is no evidence of manipulation under this assumption, there is little reason to believe observers manipulated LOE-Cont under realistic conditions.

The manipulation test assumes observers are prescient, removes achievement and growth scores from LOE-Cont creating what I characterize as *prescient* LOE-Cont, then tests for manipulation using prescient LOE-Cont (see Supplementary Appendix C for details). A *prescient* LOE-Cont score of zero means that the observer generates the exact observation score placing the teacher at a threshold. I characterize this version of LOE-Cont as prescient because for an observer to do this they must have known the teacher's achievement and growth scores, which is implausible.

The robust-bias correction approach tests for manipulation at *prescient* LOE-cont values of zero (Cattaneo et al., 2016). The robust-bias corrected approach does not reject the null hypothesis of no manipulation at the 5% level at the 200, 275, 350, or 425 thresholds. Findings are insensitive to the use of triangular or Epanechnikov kernel functions. Figure 2 presents two graphs, one for the probability density function of *prescient* LOE-Cont centered at 200 and another for the function centered at 425. The left graph shows that the probability density functions (i.e., solid black lines) that approached the 200 threshold from the left and right are statistically indistinguishable (i.e., gray 95% confidence intervals overlap). Similar results are found at the 425 threshold. Graphs of *prescient* LOE-Cont centered at 275 and 350 are in Supplementary Appendix C. Because there is no evidence of manipulation under conditions of observer prescience, I conclude that manipulation under realistic conditions is implausible.

Balance Tests of Covariates Measured at Baseline

Another test validating the assumption of local randomization examines if preexisting characteristics balance at each threshold. If the process of local randomization holds,

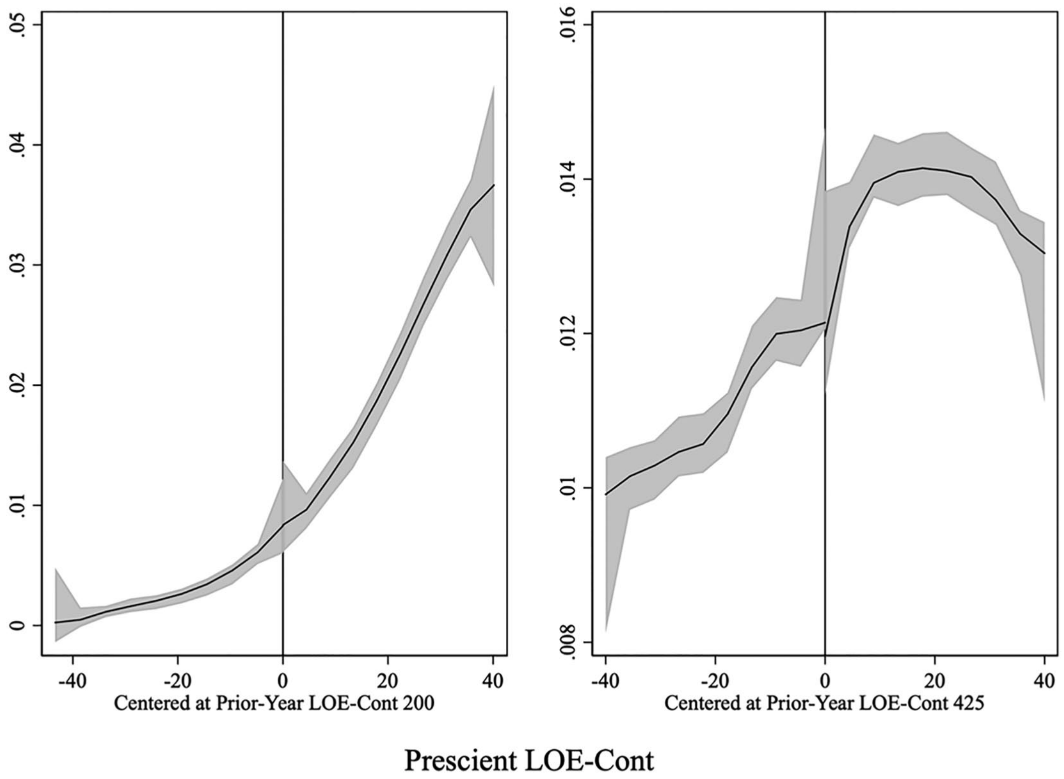

FIGURE 2. *Tests for manipulation of prescient LOE-Cont.*
Note. On each scale, a value of zero meant the observer generated the exact observation score needed to place the teacher at a prior-year LOE-Cont score of 200 or 425. Second-order local polynomials used to construct density point estimators. Epanechnikov kernel functions; 95% confidence intervals. There was no evidence of manipulation because confidence intervals overlapped at the *x* axis vale of 0 in each graph. LOE = level of effectiveness.

TABLE 1
Tests for Evidence of Local Randomization: Balance of Covariates Measured at Baseline

	200 threshold			425 threshold		
Covariate	*w* = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40
Yrs Exp: App	0.06 (0.87)	0.01 (0.84)	0.33 (0.82)	0.15 (0.32)	0.18 (0.25)	0.05 (0.21)
Yrs Exp: Prof	0.39 (1.91)	−0.74 (1.58)	−0.32 (1.37)	−0.20 (0.38)	−0.04 (0.31)	0.15 (0.28)
Female: App	−0.15 (0.10)	−0.14 (0.08)	−0.12 (0.08)	−0.03 (0.05)	0.01 (0.04)	<0.01 (0.03)
Female: Prof	−0.03 (0.10)	−0.03 (0.09)	−0.04 (0.08)	0.02 (0.02)	0.01 (0.01)	0.01 (0.01)
BA+: App	−0.03 (0.10)	−0.13 (0.08)	−0.08 (0.07)	−0.06 (0.05)	−0.05 (0.04)	−0.02 (0.04)
BA+: Prof	0.17 (0.10)	0.12 (0.09)	0.09 (0.08)	0.02 (0.02)	0.02 (0.02)	<0.01 (0.02)
Nonwhite: App	0.09 (0.07)	0.13* (0.06)	0.09 (0.06)	−0.02 (0.02)	−0.02 (0.02)	−0.02 (0.02)
Nonwhite: Prof	−0.01 (0.09)	<0.01 (0.07)	−0.02 (0.06)	−0.01 (0.01)	−0.01 (0.01)	<0.01 (0.01)
N(Tch-Yrs)	849	1,434	2,267	22,607	33,546	43,893

Note. Estimates represent the total predicted change in the outcome. Standard errors in parentheses, clustered at the teacher level. Ordinary least squares estimator employed to estimate all coefficients. BA+ is a binary variable indicating if a teacher reported earning a degree higher than a BA/BS. Nonwhite is an indicator signaling whether the teacher reported her ethnicity/ race as non-White or White.
**p* < .05.

there should not be any significant differences between the groups of teachers just to either side of any threshold. Balance tests of all baseline covariates in the vector $\boldsymbol{X}_{it}$ are conducted, regressing each baseline covariate in $\boldsymbol{X}_{it}$ on the remaining right-hand side variables from Equation 1.

There is no evidence of systematic discontinuities in baseline covariates at any of the LOE-Cont thresholds. The left and right panels of Table 1 display results from covariate balance tests at the 200 and 425 thresholds, respectively. Results from tests at the 275 and 350 thresholds are

TABLE 2

Effects of Assigning Teachers More Observations and to a Lower Prior-Year Discrete LOE on First Observation Scores

	200 threshold			425 threshold		
Certification status	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
Apprentice	*0.06 (0.11)*	*0.15 (0.10)*	*0.10 (0.09)*	−0.09* (0.04)	−0.08** (0.03)	−0.10*** (0.03)
Professional	−0.15 (0.10)	−0.10 (0.08)	−0.11 (0.07)	−0.04* (0.02)	−0.02 (0.01)	−0.01 (0.01)
N(Tch-Yrs)	1,282	2,116	3,296	22,607	33,546	43,893
% of TEAM teachers in bandwidth	1	2	3	22	32	42

Note. Italicized estimates are not associated with a change in the number of observations assigned by state policy. Apprentice teachers just below and just above the 200 threshold are both assigned four observations. Standard errors clustered at teacher level. Each model controls for teacher demographics, prior-year LOE-Cont, month of first observation, domains scored on first observation, and year fixed effects. LOE = level of effectiveness; TEAM = Tennessee Educator Acceleration Model.

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

in supplementary Appendix C. At the 200 threshold, in a bandwidth of 30, the probability that an Apprentice teacher below the threshold is non-White is significantly higher ($p < .05$) than the probability an Apprentice teacher above the threshold is non-White. There is no evidence of any imbalance among Professional teachers surrounding the 200 threshold or among Apprentice or Professional teachers surrounding the 425 threshold. Nor is there evidence of any imbalance among Apprentice teachers surrounding the 275 or 350 thresholds (see Supplementary Appendix C). Balance tests found that 275-Professional teachers in a bandwidth of 40 below the threshold are significantly more likely ($p < .05$) to hold more than a bachelor's degree and significantly less likely ($p < .05$) to be female (see Supplementary Appendix C). A total of three imbalances are detected across 96 tests (i.e., four thresholds times three bandwidths times four covariates times two levels of certification), less than expected by chance at a Type I error rate of 5%.

Results

Results from the prior-year LOE-Cont 200 and 425 thresholds are discussed first. Results from the 425 threshold and 200-Professional results capture effects from assigning more observations and a lower prior-year discrete LOE (Table 2). Apprentice teachers just to either side of the 200 threshold are not assigned different numbers of observations, but results for this group of teachers are discussed in the first section because 200-Apprentice and 200-Professional results are estimated by the same equation. Supplementary Appendix C contains descriptions of each sample at each threshold in a bandwidth of 40.

Assignment to a Lower Level of Effectiveness or More Observations

Figure 3 graphs first scores against a polynomial of prior-year LOE-Cont centered at 200 by certification status. The graphed lines represent regressions of first scores on prior-year LOE-Cont. The right panel of Figure 3 shows that Professional teachers just below the 200 threshold receive a lower first score than Professionals just above. However, Apprentice teachers just below the 200 threshold receive a higher first score than Apprentices above the threshold. Figure 4 is another binned scatterplot and regression of first scores against the forcing variable, with prior-year LOE-Cont centered at 425. Teachers below the 425 threshold receive lower first scores than teachers above the threshold.

Table 3 presents results generated by Equation 1 using data from the 200 or 425 thresholds. RDD estimates from the 200 threshold (left panel, Panel A) show that crossing from above to below the threshold raises Apprentice first scores by about 0.10 units (0.20 *SD*), but lowers the first scores of Professionals by about 0.11 units (0.22 *SD*). However, none of the estimates at the 200 threshold are significant at conventional levels. The left panel of Panel A in Table 3 shows that crossing from above to below the 425 threshold reduces the first scores of both Apprentice and Professionals. The effects on Professional teachers range from −0.01 to −0.04 (−0.02 to −0.08 *SD*) but are significant in one bandwidth only; effects on Apprentices range from −.08 to −.10 (−0.16 to −0.20 *SD*) and are significant in each bandwidth. Given the imprecision of estimates at the 200 threshold, remaining discussions in this section focus on estimates from the 425 threshold.

Sensitivity Tests. Sensitivity tests control for (a) specification of the running variable, (b) unobserved between-school differences, and (c) clustering of standard errors at the school level. Additional sensitivity tests examine the extent to which (d) teacher job satisfaction and (e) improvement efforts account for the results, and (f) explore if crossing the threshold induces teacher re-assignment to a different subject, which might explain the negative effects at the 425 threshold (see Supplementary Appendix D). No evidence suggests that any of these alternative explanations account for the original findings.

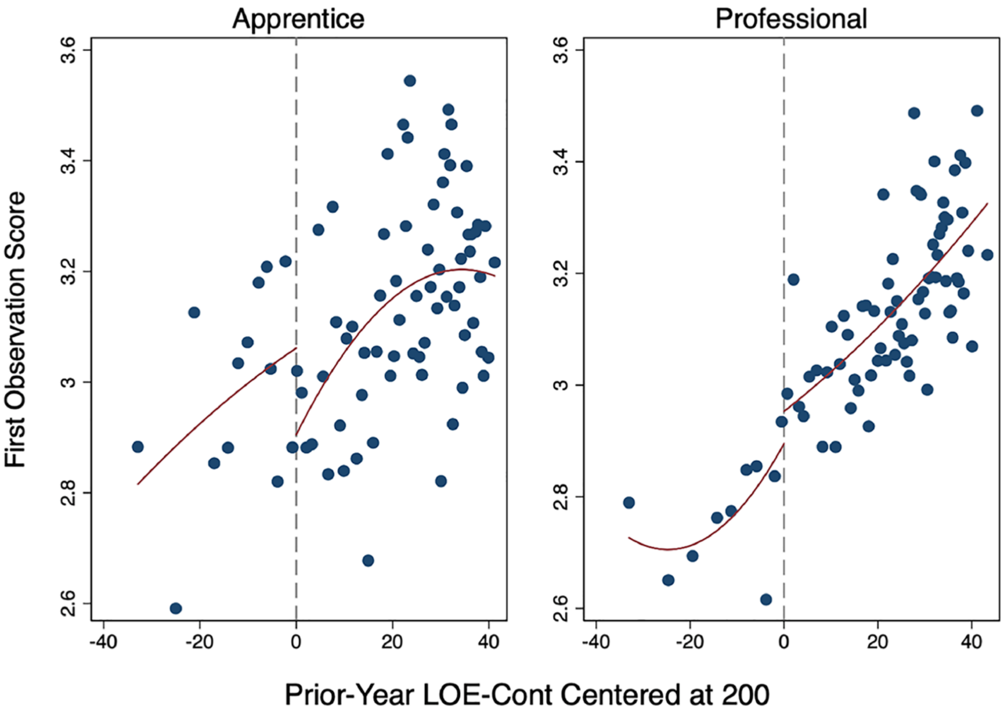

FIGURE 3. *Binned scatterplot and regression: first observation score vs. prior-year LOE-Cont at 200 threshold.*
Note. Curves are second-order polynomials of binned sample means tracing out regression of first observation scores on running variable. No control variables. *N*(App) = 1,029. *N*(Prof) = 2,267. LOE = level of effectiveness.

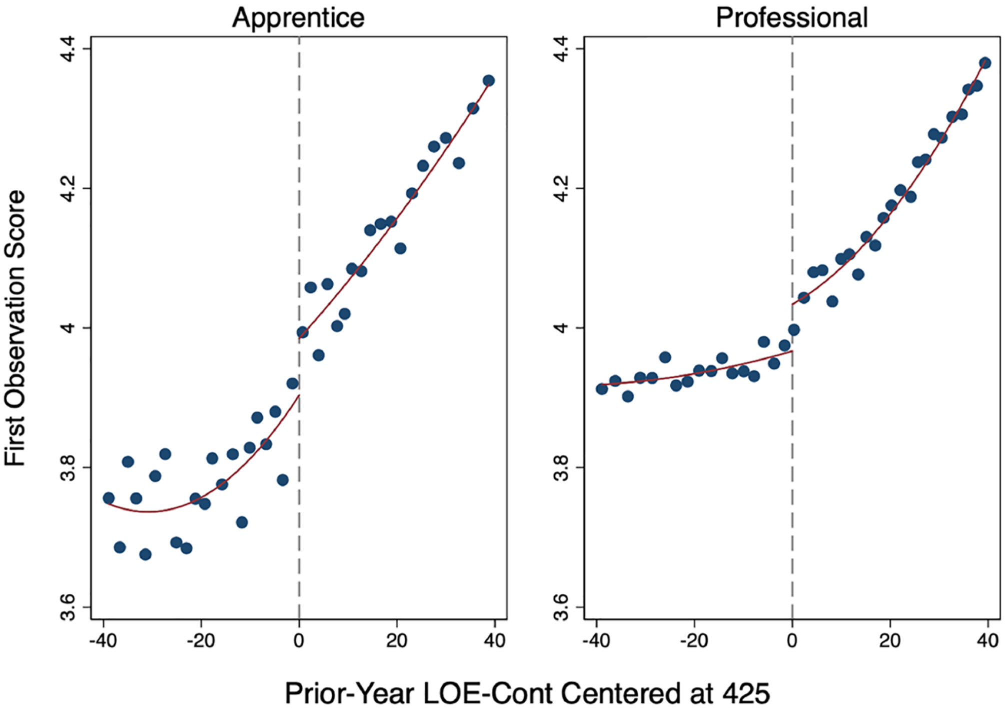

FIGURE 4. *Binned scatterplots and regressions: first observation score vs. prior-year LOE-Cont at 425 threshold by certification status.*
Note. Curves are second-order polynomials of binned sample means tracing out regression of first observation scores on running variable. No control variables. *N*(App) = 5,319. *N*(Prof) = 38,584. LOE = level of effectiveness.

TABLE 3

Moderation Analysis: RDD Effects of Assigning Apprentice Teachers More Observations and to a Lower Prior-Year Discrete LOE on First Observation Scores

	425 threshold		
Treatment and interactions	*w* = 20	*w* = 30	*w* = 40
Main effect			
Assigned More Obs	−0.14**(0.04)	−0.13***(0.04)	−0.15***(0.03)
Interactions			
Assigned More Obs × 2 Yrs Exp	0.11**(0.04)	012***(0.04)	0.11***(0.03)
Assigned More Obs × 3 Yrs Exp	0.17**(0.05)	0.17***(0.04)	0.14***(0.04)
N(Tch-Yrs)	2,935	4,193	5,319

Note. Standard errors clustered at teacher level. The reference category is Apprentice teachers with one year of experience. Each model controls for teacher demographics, LOE-Cont, discrete LOE, month of first observation, domains scored on first observation, and year fixed effects. RDD = regression discontinuity research design; LOE = level of effectiveness.
**p* < .05. ***p* < .01. ****p* < .001.

TABLE 4

Associations Between Certification Status and First Scores

Panel A. 3 or 4 years of experience			
App vs. Prof	−0.04** (0.01)	−0.03** (0.01)	*−0.01 (0.02)*
N(Tch-Yrs)	6,636	6,636	3,083
Panel B. 2 or 3 years of experience			
App vs. Prof	*0.02 [−0.01, 0.06]*	*0.01[−0.02, 0.05]*	
N(Tch-Yrs)	13,036	13,036	
Panel C. 4 or 5 years of experience			
App vs. Prof	*001 [−0.02, 0.03]*	*<0.01 [−0.02, 0.02]*	
N(Tch-Yrs)	6,271	6,271	
Prior-year LOE	LOE2–LOE4	LOE2–LOE4	LOE1 and LOE5
School fixed effects		×	

Note. Italicized estimates are not associated with a change in the number of observations assigned by state policy. Apprentice and Professional teachers within prior-year discrete LOE1 are both assigned four observations. Apprentice and Professional teachers within prior-year discrete LOE5 are both assigned one observation. Standard errors clustered at teacher level in columns one and two; clustered at school level in school fixed effects models. Each model controls for teacher demographics, prior-year LOE-Cont, prior-year discrete LOE, month of first observation, domains scored on first observation, and year fixed effects. The predictor of interest is having 3 years of experience and holding Apprentice status instead of having 4 years of experience and holding Professional status. LOE = level of effectiveness.
**p* < .05. ***p* < .01.

Moderation by Years of Experience. If the original effects at the 425 threshold are caused by assimilation bias, previously discussed research suggests that those effects may be moderated by teacher years of experience. Indeed, Apprentice teacher point estimates are always larger than Professional teacher coefficients (see Table 3). However, previously discussed research also implies that the negative effects among early-career Apprentice teachers will be most negative for teachers with the fewest years of experience. I test this implication by applying years of experience as a moderator. Specifically, a categorical experience variable is created and interacted with the variable indicating whether an Apprentice is just above or below the 425 threshold. Interaction terms represent the difference between the effects for second-year teachers and teachers with (a) 3 years or (b) 4 years of experience. Each moderated estimate compares Apprentice teachers with the same years of experience, but who were just to either side of the 425 threshold.

Results corroborate the argument that assimilation bias drove the negative effects on first scores at the 425 threshold. Second-year Apprentices just below the threshold have first scores that are about 0.14 units (0.28 *SD*) lower than second-year Apprentices just above the threshold (see Table 4). The interactions in Table 4 show that the negative effects on Apprentice first scores attenuate as early-career teachers gain experience.

Assigned More Observations, not to a Lower Level of Effectiveness

Apprentice teachers within LOE2, LOE3, and LOE4 are assigned four observations while Professional teachers are

TABLE 5
RDD Effects of Crossing Prior-Year LOE-Cont Thresholds on First Observation Scores by Certification Status

	275 threshold			350 threshold		
Certification status	w = 20	w = 30	w = 40	w = 20	w = 30	w = 40
Apprentice	0.08 (0.05)	0.04 (0.04)	<0.01 (0.02)	<−0.01 (0.04)	−0.01 (0.02)	0.02 (0.03)
Professional	0.03 (0.03)	0.02 (0.02)	−0.01 (0.02)	−0.01 (0.01)	−0.02 (0.02)	−0.03 (0.01)
N(Tch-Yrs)	9,715	14,497	18,947	15,705	23,743	31,579
% of TEAM teachers in bandwidth	9	14	18	15	23	30

Note. Standard errors clustered at teacher level. Each model controls for teacher demographics, LOE-Cont, month of first observation, domains scored on first observation, and year fixed effects. The predictor of interest is crossing from above to below the 275 or 350 thresholds, where there are no discontinuities in assigned observations, for teachers in bandwidth *w*. RDD = regression discontinuity research design; LOE = level of effectiveness; TEAM = Tennessee Educator Acceleration Model.
$^*p < .05$.

assigned two (see Figure 1). Equation 2 compares Apprentice teachers with three full years of experience to Professional teachers with 4 full years, only comparing teacher groups holding the same prior-year discrete LOE. The first score of teachers assigned four observations is 0.04 (0.08 *SD*) lower than teachers assigned two (Panel A, Table 4). The association remains relatively unchanged after controlling for school fixed effects (see Table 4).

Falsification Tests. Within LOE1 and LOE5, Apprentice and Professional teachers are assigned the same number of observations. Similarly, within LOE2–LOE4, there is no discontinuity in assigned observations between teachers with 2 or 3 years of experience, or between teachers with 4 or 5 years. If the difference in first scores between Apprentice and Professional teachers within LOE2 to LOE4 is driven by years of experience and not the difference in assigned observations, then (a) Apprentice LOE1 and LOE5s will have lower first scores than Professionals within these two discrete LOE, (b) LOE2 to LOE4 teachers with 2 instead of 3 years' experience will have lower first scores, and (c) LOE2 to LOE4 teachers with 4 instead of 5 years of experience will have lower first scores. Panels A, B, and C in Table 4 show that the associations of (a), (b), and (c) are near-zero nulls. However, no result from any falsification test in Table 4 is statistically different from the original associations of −.04 or −.03 in Panel A.

Assignment to a Lower Level of Effectiveness, Not to More Observations

Teachers just to either side of the 275 and 350 thresholds are assigned to different discrete LOE, but not to different numbers of observations (see Figure 1), providing an opportunity to estimate the effects of assigning a lower discrete LOE but not more observations.

There is no evidence of assimilation bias at the 275 or 350 thresholds (Table 5). Crossing from above to below the 275 threshold raises Apprentice (left panel Table 5) first scores by about 0.04 units (0.08 *SD*), but leaves the first scores of Professional teachers relatively unchanged. However, no estimates at the 275 threshold are significant at conventional levels. The right panel of Table 5 shows that crossing from above to below the 375 threshold has near-zero effects on both Apprentice and Professional teachers. These models are reestimated using school fixed effects and produce qualitatively similar results (see Supplementary Appendix D).

Conclusions

This study draws on the theory of assimilation bias and hypothesizes that assigning teachers to a lower effectiveness score or assigning teachers to receive more observations negatively affects observation scores via assimilation bias. Either is a potential source of assimilation bias because either could hypothetically affect observer expectations of teacher performance, independent of subsequent observations. These hypotheses were examined by exploiting discontinuities in Tennessee policy assigning teachers to different levels of effectiveness scores and assigning teachers to receive different numbers of observations. There are cases when state policy assigns teachers to a lower effectiveness score and to more observations; cases where teachers are assigned more observations but not to a lower effectiveness score; and cases assigning teaches to a lower effectiveness score but not to more observations.

There is evidence of assimilation bias, but only in cases where teachers are assigned more observations. RDDs find that teachers assigned more observations are predicted to receive lower first observation scores, though some predictions are estimated imprecisely. As suggested by previous research, the strongest negative effects (−0.30 *SD*) apply to beginning teachers (i.e., early-career 425-Apprentices). To place the degree of bias in context, observation scores are predicted to decline by approximately 0.10 *SD* if the proportion of teachers' students who are black increases by 25% (Campbell & Ronfeldt, 2018). In terms of the characteristics of Tennessee teachers, 0.30 *SD* is more than half the difference in raw average observation scores between first- and second-year teachers. The magnitude of assimilation bias suggests that the study findings are not just of theoretical importance. The degree of assimilation bias is substantially larger than other sources of bias found in previous studies. Ancillary analyses explore alternative explanations for the strong negative effects at the 425 threshold, but no evidence supports any alternative explanation.

Although evidence from cases assigning teachers to more observations and lower effective scores suggested negative assimilation bias, findings from these cases cannot disentangle the source of bias. However, results from other cases suggest that the source may be teacher assignment to more observations. Local ordinary least squares and school fixed effects regressions compared Apprentices who are assigned four observations to Professionals assigned two, but who share the same prior-year effectiveness scores. These associations suggest that assigning teachers more observations introduces assimilation bias. Tests attempting to falsify this conclusion are unable to do so, again implying that the assignment of observations drives assimilation bias. Evidence generated by comparisons of teachers assigned to different effective scores but not to different number of observations further corroborate this conclusion.

Limitations

There are three potential limitations to this study. First, the findings may not generalize to other observation systems. Previous work suggests that observations can be classified according to several qualities, including scoring procedures and observer training (Bell et al., 2019). The effects of assimilation bias may vary by any of these qualities. In the study context, observers received a total of two to four days of summer training (Alexander, 2016). Observers in other systems may receive ongoing support from support providers, which previous research suggests may reduce some forms of bias (Congdon & McQueen, 2000). Future work should examine the extent to which assimilation bias might vary by the characteristics of observer training and supports.

Relatedly, more research should examine bias in field settings. We know a great deal about the sources of bias in experimental observation scores generated by trained researcher-observers from the MET project (e.g., Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). However, it is unclear if practitioner-generated scores are susceptible to similar sources of bias. Presumably, MET researchers and subjects did not have a history of potentially bias-attenuating interactions.

This study is also limited in that it cannot identify why assimilation bias seems to be driven by the assignment of observations and not teacher effectiveness scores. At face value, one may think that observers would be more sensitive to the latter. However, the Tennessee teacher effectiveness scores are partially based on value-added scores, a measure that principals tend to view with great skepticism (Collins, 2014; Goldring et al., 2015; Hallinger et al., 2014; J. Murphy et al., 2013). Because many Tennessee teacher effectiveness scores, including the effectiveness scores of most “untested” teachers, are a function of value-added scores, observers may be skeptical of the information captured by these measures, which may explain why effectiveness scores do not affect assimilation bias. Future research might explore how observers make sense of information produced by next-generation teacher evaluation systems and how that information might introduce bias into observation scores. Interviews, surveys, and other qualitative methods may be well-suited for such investigations. Qualitative research may also be able to explore the extent to which observers consciously or unconsciously engage in assimilation bias.

Third, schools with multiple observers may endogenously sort observers to teachers. Although several sensitivity tests suggest that such within-school sorting does not explain the main results, the evidence from these tests is relatively weak. Future work should explore patterns of observer sorting; observer fixed effects are well-suited for such explorations.

Implications

Many teachers work in states that attach high-stakes consequences to teacher evaluation, and scores produced by these systems tend to be most influenced by observation scores (American Institutes for Research, 2016; Cohen & Goldhaber, 2016; National Council on Teacher Quality, 2019b). Additionally, several states assign observations based on prior-year teacher performance or productivity (American Institutes for Research, 2016; National Council on Teacher Quality, 2019a). In conjunction with the magnitude of assimilation bias, these conditions suggest policymakers or practitioners might consider how to reduce the effects and 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 and prior-year effectiveness scores. If policy makers move toward this approach, it underscores the need for additional research examining other sources of bias in observation scores generated in field settings, and research examining quantitative methods capable of removing bias in observation scores.

If post-observation feedback is influenced by assimilation bias, the effectiveness of observations as a tool for teacher development may be inhibited. 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 would misallocate professional development resources and teacher time, implying that policy makers and education agencies might consider addressing the root causes of assimilation bias.

One way to reduce assimilation bias arising from the assignment of observations by state policy is to assign teachers the same number of observations. In this study, the strongest evidence of assimilation bias existed among early-career teachers assigned four observations instead of one. Tennessee policymakers could assign all early-career teachers four observations, removing the seeming driver of assimilation bias. However, assigning all early-career teachers four observations instead of one would increase the administrative burdens of teacher evaluation, which school administrators report is already time intensive (Kraft & Gilmour, 2016; Rigby, 2015). Ideally, the decision to increase observations should at least weigh the costs of assimilation bias against the costs, broadly defined, of increasing evaluation-related administrative burdens. However, little, if any, research quantifies the administrative burdens of teacher evaluation in terms of administrator or teacher outcomes. In the absence of such research, it is unclear whether increasing the number of observations assigned to early-career teachers represents a net benefit or cost.

Additional observer professional development may represent a second way to mitigate the cause of assimilation bias. However, if the effectiveness of annual observer training resembles the effectiveness of typical annual teacher trainings, educators should not expect assimilation bias to mitigate substantially; research finds that effective professional development occurs frequently and is tailored to school- or district-specific needs (Desimone et al., 2002; Garet et al., 2001; Penuel et al., 2007). Once-per-year annual observer workshops would almost certainly not be frequent or specific enough to address contextual needs and substantially mitigate assimilation bias. Additionally, increasing observer professional development would increase the cost of teacher observation systems. Policymakers will need to decide if the costs associated with ongoing observer professional development outweigh the costs of assimilation bias identified in this study.

Acknowledgments

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

ORCID iD

Seth B. Hunter https://orcid.org/0000-0002-3051-872X

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 et al., 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 et al., 2015).

[4] Aside from the assignment of "tenure" at the 350 threshold, the only other state policies triggered by crossing an LOE-Cont threshold.

[5] Currently, it is unclear whether or not Tennessee administrative data identify who conducts observations. Tennessee administrative data contains a variable ostensibly identifying who conducted the observations. However, this variable identifies who entered the observation scores into the state information management system, who may differ from the observer. Indeed, about 100 of 1,000 Tennessee school administrators who responded to a recent statewide survey reported that they do not enter the scores that they conducted into the state system (Periscopic, 2019). Taking the observer-identifier variable at face value, 38% of all observers are principals, 33% assistant principals, 16% full-time teacher evaluators, and the remainder are district personnel.

[6] The typical teacher received two observations, and the typical teacher-respondent to the Tennessee Department of Education Educator Survey reported spending a total of one to two hours in observations each year (Tennessee Department of Education, n.d.). Dividing 1 hour by two observations results in 30 minutes per observation.

[7] Previous research finds observers tend to generate more accurate scores in the environmental domain of teaching, relative to the instructional domain (Cash et al., 2012).

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

[9] A sensitivity test also estimates school-clustered standard errors, which did not alter the statistical significance of any results (Supplementary Table D5).

[10] As discussed in the Study Context section, student outcomes are achievement or growth scores. Each score is an integer. The polychoric correlation between prior-year and contemporaneous achievement scores was 0.37, and the correlation between growth scores was 0.50.

References

Alexander, K. (2016). *TEAM administrator evaluator training*. Tennessee Department of Education. http://team-tn.org/wp-content/uploads/2013/08/TEAM-Admin-Training-2016_FINAL_PDF.pdf

Almy, S. (2011). *Fair to everyone: Building the balanced teacher evaluations that educators and students deserve*. Education Trust. https://edtrust.org/wp-content/uploads/2013/10/Fair_To_Everyone_0.pdf

American Institutes for Research. (2016). Databases on state teacher and principal evaluation policies (STEP Database and SPEP Database). http://resource.tqsource.org/stateevaldb/Compare50States.aspx

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, *37*(2), 65–75. https://doi.org/10.3102/0013189X08316420

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, *49*, 141–168. https://doi.org/10.1146/annurev.psych.49.1.141

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

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

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-3456-0

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*, *55*(6), 1233–1267. https://doi.org/10.3102/0002831218776216

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *Journal of School Psychology*, *44*(6), 473–490. https://doi.org/10.1016/j.jsp.2006.09.001

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), 378–387. https://doi.org/10.3102/0013189X16659442

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS Education Value-Added Assessment System (EVAAS®). *Education Policy Analysis Archives*, *22*(98). https://doi.org/10.14507/epaa.v22.1594

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*(2), 163–178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management*, *22*(1), 139–162. https://doi.org/10.1177/014920639602200106

Daley, G., & Kim, L. (2010). *A teacher evaluation system that works* [Working Paper]. National Institute for Excellence in Teaching.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24*(2), 81–112. https://doi.org/10.3102/01623737024002081

Donaldson, M. L. (2012). *Teachers' perspectives on evaluation reform*. Center for American Progress.

Donaldson, M. L., & Papay, J. P. (2014). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (pp. 190–209). Routledge.

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

Evans, L. (2001). Delving deeper into morale, job satisfaction and motivation among education professionals: Re-Examining the leadership dimension. *Educational Management & Administration*, *29*(3), 291–306. https://doi.org/10.1177/0263211X010293004

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945. https://doi.org/10.3102/00028312038004915

Georgia Department of Education. (2012). *Teacher keys and leader keys effective systems*. http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/Pilot Report 12-13-2012 FINAL Clean.pdf

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96–104. https://doi.org/10.3102/0013189X15575031

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform, U.S. Department of Education.

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

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445–470. https://doi.org/10.1086/669901

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, *26*(1), 5–28. https://doi.org/10.1007/s11092-013-9179-5

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

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (pp. 1–34). Bill & Melinda Gates Foundation. https://k12education.gatesfoundation.org/download/?Num=2520&filename=MET_Reliability-of-Classroom-Observations_Research-Paper.pdf

Hogan, E. A. (1987). Effects of prior expectations on performance ratings: A longitudinal study. *Academy of Management Journal*, *30*(2), 354–368. https://doi.org/10.5465/256279

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *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

Kingstrom, P. O., & Mainstone, L. E. (1985). An investigation of the rater-ratee acquaintance and rater bias. *Academy of Management Journal*, *28*(3), 641–653. https://doi.org/10.5465/256119

Koedel, C., Li, J., Springer, M. G., & Tan, L. (2015). *Do evaluation ratings affect teachers' professional development activities?* https://pdfs.semanticscholar.org/25b5/cf7e11f421aff31bcc31d8565cd51bb16d70.pdf

Koedel, C., Li, J., Springer, M. G., & Tan, L. (2017). The impact of performance ratings on job satisfaction for public school teachers. *American Educational Research Journal*, *54*(2), 241–278. https://doi.org/10.3102/0002831216687531

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

Lane, J. L. (2019). Maintaining the frame: Using frame analysis to explain teacher evaluation policy implementation. *American Educational Research Journal*, *57*(1), 5–42. https://doi.org/10.3102/0002831219848689

Lawler, E. E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, *51*(5, Pt.1), 369–381. https://doi.org/10.1037/h0025095

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, *122*(1), 159–208. https://doi.org/10.1162/qjec.122.1.159

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), 147–156. https://doi.org/10.1037/0021-9010.69.1.147

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.). Jossey-Bass. https://doi.org/10.1002/9781119210856.ch2

Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation: The case of the missing clothes? *Educational Researcher*, *42*(6), 349–354. https://doi.org/10.3102/0013189X13499625

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage.

Murphy, K. R., & Deshon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*(4), 873–900. https://doi.org/10.1111/j.1744-6570.2000.tb02421.x

National Council on Teacher Quality. (2019a). *Frequency of evaluation and observation national results* [Data set]. State Teacher Policy Database. https://www.nctq.org/yearbook/national/Frequency-of-Evaluation-and-Observation-95

National Council on Teacher Quality. (2019b). *Yearbook: State teacher policy database*. https://www.nctq.org/yearbook/home

Ost, B. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, *6*(2), 127–151. https://doi.org/10.1257/app.6.2.127

Ost, B., & Schiman, J. C. (2015). Grade-specific experience, grade reassignments, and teacher turnover. *Economics of Education Review*, *46*, 112–126. https://doi.org/10.1016/j.econedurev.2015.03.004

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

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional strategies development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, *44*(4), 921–958. https://doi.org/10.3102/0002831207308221

Periscopic. (2019). *Tennessee Educator Survey*. Tennessee Educator Survey Results. http://educatorsurvey.tnk12.gov/#1/all-districts/all-schools/0&participant=admin&result=special

Putman, H., Ross, E., & Walsh, K. (2018). *Making a difference: Six places where teacher evaluation systems are getting results*. National Council on Teacher Quality. https://www.nctq.org/dmsView/NCTQ_Report_-_Making_a_Difference

Rice, R. W., Gentile, D. A., & McFarlin, D. B. (1991). Facet importance and job satisfaction. *Journal of Applied Psychology*, *76*(1), 31–39. https://doi.org/10.1037/0021-9010.76.1.31

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

SAS. (2015). *Technical documentation for 2015 TVAAS analyses 1.1*.

Shen, J., Leslie, J. M., Spybrook, J. K., & Ma, X. (2012). Are principal background and school processes related to teacher job satisfaction? A multilevel study using schools and staffing survey 2003-04. *American Educational Research Journal*, *49*(2), 200–230. https://doi.org/10.3102/0002831211419949

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), 340–359. 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*, *38*(2), 293–317. 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), 436–442. https://doi.org/10.1037/0021-9010.81.4.436

Tennessee Department of Education. (n.d.). *Evaluation*. https://team-tn.org/teacher-evaluation

Tennessee Department of Education. (2016). *Evaluation Guidance | TEAM-TN*. Tennessee Educator Acceleration Model. http://team-tn.org/evaluation/evaluation-guidance/

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

Tennessee State Board of Education. (2013). *Teacher and Principal Evaluation Policy 5.201*. https://www.tn.gov/content/dam/tn/stateboardofeducation/documents/2013_sbe_meetings/october_25_2013_sbe_meeting/10-25-13%20III%20H%20Teacher%20and%20Principal%20Evaluation%20Policy%205%20201%20Attachment.pdf

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. https://www.nctq.org/dmsView/Final_Evaluation_Paper

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

Author

SETH B. HUNTER is an assistant professor of education leadership at George Mason University. His research interests include the effects of teacher leadership and teacher leader labor markets, and the policies and practices of educator evaluation.